面向21世纪课程教材

信息管理与信息系统专业教材系列

运筹学 第4版

Operations Research

《运筹学》教材编写组 ◎ 编

清华大学出版社

VISIT…

LANZAROTE Caliente.com

面向21世纪课程教材
信息管理与信息系统专业教材系列

管理信息系统（第6版）	薛华成
信息管理基础	岳剑波
因特网信息资源检索与利用（第2版）	符绍宏
信息组织（第3版）	马张华
信息经济学	靖继鹏
运筹学（第4版）	《运筹学》教材编写组

本书提供教学课件，教师可登录http://www.tup.com.cn下载。

清华大学出版社数字出版网站
WQBook 文泉书局
www.wqbook.com

ISBN 978-7-302-28879-4
9 787302 288794 >
定价：39.80元

面向21世纪课程教材
信息管理与信息系统专业教材系列

《运筹学》教材编写组 ◎ 编

运筹学 第4版

Operations Research

清华大学出版社
北京

内 容 简 介

本书在第三版基础上，吸收了广大读者的意见，做了局部调整和修改。全书共分为绪论、线性规划与目标规划、整数线性规划、非线性规划、动态规划、图与网络分析、排队论、存储论、对策论、决策论和启发式方法11篇。

本书着重介绍运筹学的基本原理和方法，注重结合经济管理专业实际，具有一定的深度和广度。书中每章后附有习题，便于自学。有些部分的后面增补了"注记"，便于读者了解运筹学各分支的发展趋势。

本书可作为高等院校理工科各专业的教材，亦可作为报考研究生的参考书。

本书封面贴有清华大学出版社防伪标签，无标签者不得销售。
版权所有，侵权必究。侵权举报电话：010-62782989　13701121933

图书在版编目(CIP)数据

运筹学/《运筹学》教材编写组编. --4 版. --北京：清华大学出版社，2012.9
(面向21世纪课程教材·信息管理与信息系统专业教材系列)
ISBN 978-7-302-28879-4

Ⅰ. ①运…　Ⅱ. ①运…　Ⅲ. ①运筹学－高等学校－教材　Ⅳ. ①O22

中国版本图书馆 CIP 数据核字(2012)第 105292 号

责任编辑：贺　岩
封面设计：常雪影
责任校对：王凤芝
责任印制：张雪娇

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课 件 下 载：http://www.tup.com.cn，010-62795954
印 刷 者：清华大学印刷厂
装 订 者：三河市金元印装有限公司
经　　销：全国新华书店
开　　本：185mm×230mm　**印　　张**：35.25　**插　　页**：1　**字　　数**：747 千字
版　　次：1982 年 6 月第 1 版　　2012 年 9 月第 4 版　　**印　　次**：2012 年 9 月第1次印刷
印　　数：1～15000
定　　价：39.80 元

产品编号：047508-01

编写组成员

主　编：钱颂迪

编　委：甘应爱　田　丰　李维铮（杜　纲）　李梅生（邓成樑）

陈秉正　胡运权　顾基发　郭耀煌

前言(第4版) PREFACE

2010 年 12 月 12 日在三亚,《运筹学》第三版的老作者们举行了一次聚会,出席者有钱颂迪、郭耀煌、顾基发、甘应爱、胡运权、李梅生和陈秉正。

与会者热烈认真讨论了如何修改《运筹学》第三版的议题，并获得共识：

1. 第 4 版的修改原则。大家一致认为：要保持本书已形成的风格；要保持原书框架结构和体系,书的读者对象不变；章节之间调整按有增有减办法处理；要与时俱进,要适当体现当下的管理理念等。

2. 这次修改的总要求：消除概念的、文字的、计算的、印刷的错误；更换不合适的例子；精炼文字,表述要正确,经得起推敲。

3. 这次修改以原作者为主,根据各作者健康状况,量力而行；并且议论了今后“接力棒”修改模式问题；先在第 4 版的修改时吸收个别新成员参加。

4. 李梅生邀请邓成樑负责修改他那部分书稿；杜纲负责修改李维铮那部分书稿。

5. 充实绪论部分内容；补充线性规划部分内容：建立线性规划问题模型的三个重要假设；对影子价格的经济意义作了较全面的修改等。

6. 大家回顾了本教材从 1980 年开始建立起彼此之间的合作和友谊,和获得清华大学出版社的持久的支持,以及深深怀念已故老作者李德、李维铮、郑大本所做出的贡献。

7. 感谢多年来广大读者和老师们对本教材的关心和支持,特别感谢东北大学工商管理学院张川老师,同济大学王明照老师对本教材的修改提出的具体意见。

参加本次修订的作者为：

1. 绪论	顾基发(中国科学院系统科学研究所)
2. 线性规划与目标规划	钱颂迪(南京航天管理干部学院)
	胡运权(哈尔滨工业大学)
3. 整数线性规划	李维铮、杜纲(天津大学)
4. 非线性规划	郭耀煌(西南交通大学)
5. 动态规划	甘应爱(华中科技大学)
6. 图与网络分析	
第 11 章	田丰(中国科学院系统科学研究所)
第 12 章	钱颂迪
7. 排队论	李维铮、杜纲

8. 存储论　　　　　　　　李梅生、邓成梁（华中科技大学）
9. 对策论　　　　　　　　陈秉正(清华大学)
10. 决策论
    第 16 章　　　　　　　钱颂迪、顾基发
    第 17 章　　　　　　　顾基发
11. 启发式方法　　　　　　郭耀煌

本书如有不妥之处,敬请广大读者批评指正。

钱颂迪

2012.4

前言(第1版) PREFACE

为了实现我国的四个现代化，我们不但要学习和掌握先进的科学技术，而且要学习和掌握现代化的科学管理方法。近几年来，我们从管理实践中更加认识到，由于计划和管理不当，在时间、人力、物力和资金等方面造成了很大的浪费，从而给我国的经济建设带来了严重损失。为了适应现代化管理的需要，最近几年在我国许多工科院校中，相继建立了一些工业经济、工商管理或系统工程等系或专业，并且都开设了运筹学的课程。

运筹学是近 40 年来发展起来的一门新兴学科。它的目的是为行政管理人员在做决策时提供科学的依据。因此，它是实现管理现代化的有力工具。运筹学在生产管理、工程技术、军事作战、科学试验、财政经济以及社会科学中都得到了极为广泛的应用。

应用运筹学去处理问题时，有两个重要的特点：一是从全局的观点出发；二是通过建立模型，如数学模型或模拟模型，对于需要求解的问题给出最合理的决策。在建立模型和求解的过程中，往往要用到一些数学方法和技巧。因此，许多运筹学工作者，特别是中国的运筹学工作者，往往都是来自数学专业。由于这个原因，目前国内流行的有关运筹学的教科书，多半偏重于数学方法的论证，对于解决实际问题时所需要的建立模型的概念与解题的技巧不够重视。这种情况不太适宜于工科院校学生的需要。本书是专为工科院校的经济管理专业的学生编写的。内容上力求深入浅出，文字通俗易懂，方法上着重于思路和几何的直观解释，并尽量结合经济管理专业举一些实例。

本书是在较短时间内完成的，作为第一次尝试编写一本适宜于工科院校的运筹学教科书，无疑对我国运筹学教学以及促进运筹学的研究都将是有意义的。

中国数学会运筹学会　许国志　桂湘云

1981.9

目 录 CONTENTS

第1篇 绪 论

第1章 运筹学概论 …… 3

1.1 运筹学的简史 …… 3
1.2 运筹学的性质和特点 …… 5
1.3 运筹学的工作步骤 …… 5
1.4 运筹学的模型 …… 6
1.5 运筹学的应用 …… 7
1.6 运筹学的展望 …… 10
参考资料 …… 11

第2篇 线性规划与目标规划

第2章 线性规划与单纯形法 …… 15

2.1 线性规划问题及其数学模型 …… 15
2.2 线性规划问题的几何意义 …… 23
2.3 单纯形法 …… 27
2.4 单纯形法的计算步骤 …… 37
2.5 单纯形法的进一步讨论 …… 41
2.6 应用举例 …… 47
习题 …… 55

第3章 对偶理论和灵敏度分析 …… 59

3.1 单纯形法的矩阵描述 …… 59
3.2 单纯形法的矩阵计算 …… 60
3.3 对偶问题的提出 …… 64
3.4 线性规划的对偶理论 …… 64
3.5 影子价格 …… 72
3.6 对偶单纯形法 …… 74

3.7 灵敏度分析 …… 77
3.8* 参数线性规划 …… 84
习题 …… 87

第4章 运输问题 …… 92

4.1 运输问题的数学模型 …… 92
4.2 表上作业法 …… 93
4.3 产销不平衡的运输问题及其求解方法 …… 104
4.4 应用举例 …… 107
习题 …… 113

第5章 线性目标规划 …… 116

5.1 目标规划的数学模型 …… 116
5.2 解目标规划的图解法 …… 118
5.3 解目标规划的单纯形法 …… 120
5.4 应用举例 …… 123
习题 …… 127
参考资料 …… 129

第3篇 整数线性规划

第6章 整数线性规划 …… 133

6.1 整数线性规划问题的提出 …… 133
6.2 分支定界解法 …… 134
6.3 割平面解法 …… 138
6.4 0-1型整数线性规划 …… 142
6.5 指派问题 …… 146
习题 …… 152
参考资料 …… 154

第4篇 非线性规划

第7章* 无约束问题 …… 157

7.1 基本概念 …… 157
7.2 一维搜索 …… 170

7.3　无约束极值问题的解法 …… 176

第 8 章*　约束极值问题 …… 199

8.1　最优性条件 …… 199
8.2　二次规划 …… 203
8.3　可行方向法 …… 206
8.4　制约函数法 …… 210
习题 …… 217
参考资料 …… 220

第 5 篇　动 态 规 划

第 9 章　动态规划的基本方法 …… 225

9.1　多阶段决策过程及实例 …… 225
9.2　动态规划的基本概念和基本方程 …… 226
9.3　动态规划的最优性原理和最优性定理 …… 235
9.4　动态规划和静态规划的关系 …… 237
习题 …… 246

第 10 章　动态规划应用举例 …… 249

10.1　资源分配问题 …… 249
10.2　生产与存储问题 …… 261
10.3*　背包问题 …… 271
10.4*　复合系统工作可靠性问题 …… 275
10.5　排序问题 …… 277
10.6　设备更新问题 …… 280
10.7*　货郎担问题 …… 283
习题 …… 285
参考资料 …… 290

第 6 篇　图与网络分析

第 11 章　图与网络优化 …… 293

11.1　图的基本概念 …… 294
11.2　树 …… 298

11.3 最短路问题 …… 304
11.4 网络最大流问题 …… 312
11.5 最小费用最大流问题 …… 318
11.6 中国邮递员问题 …… 321
习题 …… 325
参考资料 …… 329

第12章 网络计划 …… 330

12.1 网络计划图 …… 330
12.2 网络计划图的时间参数计算 …… 334
12.3 时标网络计划图 …… 338
12.4 网络计划的优化 …… 339
12.5 网络计划软件 …… 343
习题 …… 345
参考资料 …… 346

第7篇 排 队 论

第13章 排队论 …… 349

13.1 基本概念 …… 349
13.2 到达间隔的分布和服务时间的分布 …… 354
13.3 单服务台负指数分布排队系统的分析 …… 361
13.4 多服务台负指数分布排队系统的分析 …… 372
13.5 一般服务时间 $M/G/1$ 模型 …… 379
13.6 经济分析——系统的最优化 …… 382
13.7* 分析排队系统的随机模拟法 …… 386
习题 …… 390

第8篇 存 储 论

第14章 存储论 …… 397

14.1 存储论的基本概念 …… 397
14.2 确定性存储模型 …… 400
14.3 随机性存储模型 …… 414
14.4 其他类型存储问题 …… 428

习题 …… 434
参考资料 …… 436

第9篇 对 策 论

第15章 对策论基础 …… 439

15.1 引言 …… 439
15.2 矩阵对策的基本定理 …… 442
15.3 矩阵对策的解法 …… 455
15.4* 其他类型对策简介 …… 466
习题 …… 473
参考资料 …… 476

第10篇 决 策 论

第16章 单目标决策 …… 479

16.1 决策的分类 …… 479
16.2 决策过程 …… 480
16.3 不确定型的决策 …… 481
16.4 风险决策 …… 485
16.5 效用理论在决策中的应用 …… 491
16.6 决策树 …… 494
16.7 灵敏度分析 …… 498
习题 …… 499
参考资料 …… 502

第17章 多目标决策 …… 503

17.1 引言 …… 503
17.2 基本概念 …… 504
17.3 化多为少的方法 …… 507
17.4 分层序列法 …… 515
17.5 直解求非劣解 …… 516
17.6 多目标线性规划的解法 …… 517
17.7 层次分析法 …… 522
参考资料 …… 527

第11篇　启发式方法

第18章*　启发式方法 …… 531

18.1　基本概念 …… 531

18.2　应用及例子 …… 534

习题 …… 544

参考资料 …… 547

*表示可选讲内容

Part 1

第 1 篇

绪论

CHAPTER 1 第1章

运筹学概论

1.1 运筹学的简史

运筹学作为科学名字出现在20世纪30年代末。当时英、美对付德国的空袭，雷达作为防空系统的一部分，从技术上是可行的，但实际运用时却并不好用。为此一些科学家研究如何合理运用雷达开始进行一类新问题的研究。因为它与研究技术问题不同，就称之为“运用研究”。

运筹学的英文名词 operational research 是英国人最早在20世纪30年代末提出来的，很快美国也跟上，但用了 operations research，它们英文简写都是 OR。我国在1956年曾用过运用学的名词，到1957年正式定名为运筹学。为了进行运筹学研究，在英、美的军队中成立了一些专门小组，开展了护航舰队保护商船队的编队问题和当船队遭受德国潜艇攻击时，如何使船队损失最少的问题的研究。研究了反潜深水炸弹的合理爆炸深度后，使德国潜艇被摧毁数增加到400%；研究了船只在受敌机攻击时，提出了大船应急速转向和小船应缓慢转向的逃避方法。研究结果使船只在受敌机攻击时，中弹数由47%降到29%。当时研究和解决的问题都是短期的和战术性的。第二次世界大战后在英、美军队中相继成立了更为正式的运筹研究组织。

由于这门学问当时与军事有关，因此有些内容处于保密状态。很多年以后，到20世纪40年代末50年代初，其中一些与军事密切相关的内容才逐渐公开出来，其中莫尔斯(P. M. Morse)与金博尔(G. E. Kimball)1951年出版的《运筹学方法》(*Methods of Operations Research*)一书可以作为那段时期重要工作的总结。后来以兰德公司(RAND)为首的一些部门开始着重研究战略性问题、未来的武器系统的设计和其可能合理运用的方法。例如为美国空军评价各种轰炸机系统，讨论了未来的武器系统和未来战争的战略。他们还研究了苏联的军事能力及未来的预报，分析苏联政治局计划的行动原则和将来的行动预测。到20世纪50年代，由于开发了各种洲际导弹，到底发展哪种导弹，运筹学界也参与了争论。兰德公司接着提出了系统分析(systems analysis, SA)的名词及其相应技术和方法，其应用开始更偏重于战略方面。他们参与了战略力量的构成和数量问题研究。除军事方面的应用研究以外，相继在工业、农业、经济和社会问题等各领域都有应用，后来也有人把这两个词放在一起叫 SA/OR。与此同时，运筹数学有了飞快

的发展,并形成了运筹学的许多分支。如数学规划(线性规划、非线性规则、整数规划、目标规划、动态规划、随机规划等)、图论与网络、排队论(随机服务系统理论)、存储论、对策论、决策论、维修更新理论、搜索论、可靠性和质量管理等。作为运筹学的早期工作其历史可追溯到1914年,军事运筹学中的兰彻斯特(Lanchester)战斗方程是在1914年提出的。排队论的先驱者丹麦工程师爱尔朗(Erlang)1917年在哥本哈根电话公司研究电话通信系统时,提出了排队论的一些著名公式。存储论的最优批量公式是在20世纪20年代初提出的。在商业方面列温逊在20世纪30年代已用运筹思想分析商业广告、顾客心理。线性规划是由丹捷格(G. B. Dantzig)在1947年发表的成果。所解决的问题是美国制定空军军事规划时提出的,并提出了求解线性规划问题的单纯形法。而早在1939年苏联学者康托洛维奇(Л. В. Канторович)在解决工业生产组织和计划问题时,已提出了类似线性规划的模型,并给出了"解乘数法"的求解方法。由于当时未被领导重视,直到1960年康托洛维奇再次发表了《最佳资源利用的经济计算》一书后,才受到国内外的一致重视。为此康托洛维奇得到了诺贝尔奖。值得一提的是丹捷格认为线性规划模型的提出是受到了列昂节夫的投入产出模型(1932年)的影响;后来列昂节夫的投入产出也得到了诺贝尔奖。关于线性规划的理论是受到了冯·诺依曼(Von Neumann)的帮助。冯·诺依曼和摩根斯坦(O. Morgenstern)合著的《对策论与经济行为》(1944年)是对策论的奠基作,同时该书已隐约地指出了对策论与线性规划对偶理论的紧密联系。线性规划提出后很快受到经济学家的重视,如在第二次世界大战中从事运输模型研究的美国经济学家库普曼斯(T. C. Koopmans),他很快看到了线性规划在经济中应用的意义,并呼吁年轻的经济学家要关注线性规划。库普曼斯在1975年获诺贝尔经济奖。其中阿罗、萨谬尔逊、西蒙、多夫曼和胡尔威茨等都获得了诺贝尔奖,并在运筹学某些领域中发挥过重要作用。我们初步统计了到2007年为止共有19个诺贝尔奖获得者的研究与运筹学有关。回顾一下最早投入运筹学领域工作的诺贝尔奖获得者、美国物理学家勃拉凯特(Blackett)领导的第一个以运筹学命名的小组是有意义的。由于该小组的成员复杂,人们戏称它为勃拉凯特马戏团,其实是一个由各方面专家组成的交叉学科小组。从以上简史可见,为运筹学的建立和发展作出贡献的有物理学家、经济学家、数学家、其他专业的学者、军官和各行业的实际工作者。

最早建立运筹学会的国家是英国(1948年),接着是美国(1952年)、法国(1956年)、日本和印度(1957年)等。到2005年为止,国际上已有48个国家和地区建立了运筹学会或类似的组织。我国的运筹学会成立在1980年。1959年由英、美、法三国的运筹学会发起成立了国际运筹学联合会(IFORS),以后各国的运筹学会纷纷加入,我国于1982年加入该会。此外还有一些地区性组织,如欧洲运筹学协会(EURO)成立于1975年,亚太运筹学协会(APORS)成立于1985年。

20世纪50年代中期,钱学森、许国志等教授将运筹学由西方引入我国,并结合我国的特点在国内推广应用。

他们最早在中国科学院力学所建立了运筹室，在运筹学多个领域开展研究和应用工作，其中在经济数学方面，特别是投入产出表的研究和应用开展较早。质量控制(后改为质量管理)的应用也有特色。在此期间以华罗庚教授为首的一大批数学家加入到运筹学的研究队伍，在中国科学院数学所也建立了运筹室，使运筹数学的很多分支很快跟上当时的国际水平。

1.2 运筹学的性质和特点

运筹学是一门应用科学，至今还没有统一且确切的定义。提出以下几个定义来说明运筹学的性质和特点。莫斯(P. M. Morse)和金博尔(G. E. Kimball)曾对运筹学下的定义是："为决策机构在对其控制下业务活动进行决策时，提供以数量化为基础的科学方法。"它首先强调的是科学方法，这含义不单是某种研究方法的分散和偶然的应用，而是可用于整个一类问题上，并能传授和有组织的活动。它强调以量化为基础，必然要用数学。但任何决策都包含定量和定性两方面，而定性方面又不能简单地用数学表示，如政治、社会等因素，只有综合多种因素的决策才是全面的。运筹学工作者的职责是为决策者提供可以量化方面的分析，指出那些定性因素的力度。另一定义是："运筹学是一门应用科学，它广泛应用现有的科学技术知识和数学方法，解决实际中提出的专门问题，为决策者选择最优决策提供定量依据。"这定义表明运筹学具有多学科交叉的特点，如综合运用经济学、心理学、物理学、化学中的一些方法。运筹学是强调最优决策，"最"是过分理想了，在实际生活中往往用次优、满意等概念代替最优。因此，运筹学的又一定义是："运筹学是一种给出问题坏的答案的艺术，否则的话问题的结果会更坏。"

为了有效地应用运筹学，前英国运筹学学会会长托姆林森提出六条原则：

(1) 合伙原则。是指运筹学工作者要和各方面人，尤其是同实际部门工作者合作。

(2) 催化原则。在多学科共同解决某问题时，要引导人们改变一些常规的看法。

(3) 互相渗透原则。要求多部门彼此渗透地考虑问题，而不是只局限于本部门。

(4) 独立原则。在研究问题时，不应受某人或某部门的特殊政策所左右，应独立从事工作。

(5) 宽容原则。解决问题的思路要宽，方法要多，而不是局限于某种特定的方法。

(6) 平衡原则。要考虑各种矛盾的平衡，关系的平衡。

1.3 运筹学的工作步骤

运筹学在解决大量实际问题过程中形成了自己的工作步骤。

(1) 提出和形成问题。即要弄清问题的目标，可能的约束，问题的可控变量以及有关参数，搜集有关资料；

(2) 建立模型。即把问题中可控变量、参数和目标与约束之间的关系用一定的模型表示出来；

(3) 求解。用各种手段(主要是数学方法,也可用其他方法)将模型求解。解可以是最优解、次优解、满意解。复杂模型的求解需用计算机,解的精度要求可由决策者提出；

(4) 解的检验。首先检查求解步骤和程序有无错误,然后检查解是否反映现实问题；

(5) 解的控制。通过控制解的变化过程决定对解是否要作一定的改变；

(6) 解的实施。是指将解用到实际中必须考虑到实施的问题,如向实际部门讲清解的用法,在实施中可能产生的问题和修改。

以上过程应反复进行。

1.4 运筹学的模型

运筹学在解决问题时,按研究对象不同可构造各种不同的模型。模型是研究者对客观现实经过思维抽象后用文字、图表、符号、关系式以及实体模样描述所认识到的客观对象。模型的有关参数和关系式是较容易改变的,这样有助于问题的分析和研究。利用模型可以进行一定预测、灵敏度分析等。

阿可夫等对运筹学的模型分类、构模等有较完整的描述。

模型有三种基本形式：①形象模型；②模拟模型；③符号或数学模型。目前用得最多的是符号或数学模型。构造模型是一种创造性劳动,成功的模型往往是科学和艺术的结晶,构模的方法和思路有以下五种：

(1) 直接分析法。按研究者对问题内在机理的认识直接构造出模型。运筹学中已有不少现存的模型,如线性规划模型、投入产出模型、排队模型、存储模型、决策和对策模型等。这些模型都有很好的求解方法及求解的软件,但用这些现存的模型研究问题时,要注意不能生搬硬套。

(2) 类比法。有些问题可以用不同方法构造出模型,而这些模型的结构性质是类同的,这就可以互相类比。如物理学中的机械系统、气体动力学系统、水力学系统、热力学系统及电路系统之间就有不少彼此类同的现象。甚至有些经济系统、社会系统也可以用物理系统来类比。在分析一些经济、社会问题时,不同国家之间有时也可以找出某些类比的现象。

(3) 数据分析法。对有些问题的机理尚未了解清楚,若能搜集到与此问题密切相关的大量数据,或通过某些试验获得大量数据,这样可以运用统计分析法建模。

(4) 试验分析法。当有些问题的机理不清,又不能做大量试验来获得数据,这时只能通过做局部试验的数据加上分析来构造模型。

(5) 想定(构想)法(scenario)。当有些问题的机理不清,又缺少数据,又不能做试验来获得数据时,例如一些社会、经济、军事问题,人们只能在已有的知识、经验和某些研究

的基础上，对于将来可能发生的情况给出合乎逻辑的设想和描述。然后运用已有的方法构造模型，并不断修正完善，直至比较满意为止。在研究社会问题时，国内外有人提出人工社会的构思，与这条建模思路有相似之处，人们可以利用计算机先在人工社会中进行大量的计算机试验，然后在真实社会得到验证，或者通过人工社会获得在真实社会一时未能预知的方案和结果。

模型的一般数学形式可用下列表达式描述：

目标的评价准则 $U=f(x_i, y_j, \xi_k)$

约束条件 $g(x_i, y_j, \xi_k) \geqslant 0$

式中：x_i——可控变量；

y_j——已知参数；

ξ_k——随机因素。

目标的评价准则一般要求达到最佳(最大或最小)、适中、满意等。准则可以是单一的，也可是多个的。约束条件可以没有，也可有多个。当 g 是等式时，即为平衡条件。当模型中无随机因素时，称它为确定性模型，否则为随机模型。随机模型的评价准则可用期望值，也可用方差，还可用某种概率分布来表示。当可控变量只取离散值时，称为离散模型，否则称为连续模型。也可按使用的数学工具将模型分为代数方程模型、微分方程模型、概率统计模型、逻辑模型等。若用求解方法来命名时，有直接最优化模型、数字模拟模型、启发式模型。也有按用途来命名的，如分配模型、运输模型、更新模型、排队模型、存储模型等。还可以用研究对象来命名，如能源模型、教育模型、军事对策模型、宏观经济模型等。

1.5 运筹学的应用

在介绍运筹学的简史时，已提到了运筹学在早期的应用，主要在军事领域。第二次世界大战后运筹学的应用转向民用，这里只能对某些重要领域给予简述。

(1) 市场销售。主要应用在广告预算和媒介的选择、竞争性定价、新产品开发、销售计划的制定等方面。如美国杜邦公司在20世纪50年代起就非常重视将运筹学用于研究如何做好广告工作，产品定价和新产品的引入。通用电力公司对某些市场进行模拟研究。

(2) 生产计划。在总体计划方面主要用于总体确定生产、存储和劳动力的配合等计划，以适应波动的需求计划，用线性规划和模拟方法等。如巴基斯坦某一重型制造厂用线性规划安排生产计划，节省10%的生产费用。还可用于生产作业计划、日程表的编排等。此外，还有在合理下料、配料问题、物料管理等方面的应用。

(3) 库存管理。主要应用于多种物资库存量的管理，确定某些设备的能力或容量，如停车场的大小、新增发电设备的容量大小、电子计算机的内存量、合理的水库容量等。美国某机器制造公司应用存储论后，节省18%的费用。目前国外新动向是将库存理论与计

算机的物资管理信息系统相结合。如美国西电公司,从 1971 年起用 5 年时间建立了“西电物资管理系统”,使公司节省了大量物资存储费用和运费,而且减少了管理人员。

(4) 运输问题。这涉及空运、水运、公路运输、铁路运输、管道运输、厂内运输。空运问题涉及飞行航班和飞行机组人员服务时间安排等。为此在国际运筹学协会中设有航空组,专门研究空运中的运筹学问题。水运有船舶航运计划、港口装卸设备的配置和船到港后的运行安排。公路运输除了汽车调度计划外,还有公路网的设计和分析,市内公共汽车路线的选择和行车时刻表的安排,出租汽车的调度和停车场的设立。铁路运输方面的应用就更多了。

(5) 财政和会计。这里涉及预算、贷款、成本分析、定价、投资、证券管理、现金管理等。用得较多的方法是统计分析、数学规划、决策分析。此外还有盈亏点分析法、价值分析法等。

(6) 人事管理。这里涉及六个方面,第一是人员的获得和需求估计;第二是人才的开发,即进行教育和训练;第三是人员的分配,主要是各种指派问题;第四是各类人员的合理利用问题;第五是人才的评价,其中有如何测定一个人对组织、社会的贡献;第六是工资和津贴的确定等。

(7) 设备维修、更新和可靠性、项目选择和评价。

(8) 工程的优化设计。这在建筑、电子、光学、机械和化工等领域都有应用。

(9) 计算机和信息系统。可将运筹学用于计算机的内存分配,研究不同排队规则对磁盘工作性能的影响。有人利用整数规划寻找满足一组需求文件的寻找次序,利用图论、数学规划等方法研究计算机信息系统的自动设计。

(10) 城市管理。这里有各种紧急服务系统的设计和运用,如救火站、救护车、警车等分布点的设立。美国曾用排队论方法来确定纽约市紧急电话站的值班人数。加拿大曾研究一城市的警车的配置和负责范围,出事故后警车应走的路线等。此外有城市垃圾的清扫、搬运和处理;城市供水和污水处理系统的规划等。

值得提出的是应用方面新的动向,例如,存储理论的应用已经从车间、工厂规模转向整个从用户、零售、批发、中间运输一直到工厂生产供应,形成现在的供应链的设计、管理和应用。在武器和大型装置方面,不单研究其运用,更转向设计和规划等。此外在银行、医院、经济、运输、信息系统、电子商务和电子政务等也都有了新的应用。由于涉及面太多,我们只想用美国运筹学和管理学研究协会组织举办的埃德曼奖的几个例子供大家参考。该奖由世界著名的运筹和管理科学家埃德曼(F. Edelman)于 1971 年创立,每年评一次,先评出一批候选奖,然后选出 5～6 名提名奖,最后从中评出一个最佳奖。埃德曼奖的评选原则包括:运筹学和管理科学理论和方法的创新、应用工作对企业创造的直接经济效益,以及对社会和人类生活所做的积极贡献。

例如,2002 年有 30 多个项目进入了初评。经过严格的筛选,法国标致汽车公司、美国糖果制造巨商玛氏公司、美国大陆航空公司、德国 Rhenania catalog house、美国迅达电

梯公司，以及美国先正达农业企业得到提名，并进入最终决赛。这6个项目组在经过激烈的角逐后，于刚领导的项目小组由于在大陆航空公司等民用航空企业所创造的实实在在的经济效益，以及对社会和人们生活所做出的贡献，成为2002年度大奖的获得者。于刚提交的项目内容主要是当民航班机受到各种干扰之后，航班机组人员如何以最快的时间和最佳的组合来使航班恢复正常运营。他领导开发的实时决策支持系统已在美国的联合航空公司，大陆航空公司，西北航空公司和西南航空公司运行多年。这些系统在一系列影响航班正常运营的重大事件时起到了极大的效果，例如2000年12月和2001年3月的美国北部暴风雪，2001年休斯敦的大洪水，这些对航班运营和管理所造成的重大影响都在开发的系统的支持下，将损失降低到最低程度。特别是对于2001年“9·11事件”所造成的影响，该系统为美国大陆航空公司挽回了至少3 000万美元的损失。而其在2001年全年为该公司所创造的实际价值就超过了6 000万美元。

下面我们再介绍2008年度的由6个最终提名奖中选出该年度大奖获得者荷兰铁路及其主要成果。

所提出报奖项目：荷兰火车新时刻表。

问题：到2006年荷兰的铁路从1970年的80亿人·千米增加到154亿人·千米。在不改变时刻表的结构情况下要求计划有更多更长的火车发出。2006年荷兰铁路要求开发一个新的时刻表。

运筹学解决方法：构建了一个改进的、循环的时刻表，它每一小时可以重复。构造这样的时刻表需要求解一个复杂的组合优化问题。设计了4个求解程序：(1)CADANS求解网络时刻表问题；(2)STATIONS找出火车通过各个站时详细的路径；(3)ROSA处理铁路车辆的周转；(4)TURNI对乘员的时刻表安排。

价值：从2007年对旅客数进行全时的记录。随着最大的时刻表的改进旅客数增加了10%～15%。在规定时刻表的时间内，火车准点到达率(指误差在3分钟之内)从2006年的84.8%提高到2007年的87.0%。调查表明，公众意见很快从负面批评转向正面肯定。每年增加了4 000万欧元的利润。在现有网络下还有可能减少拥堵和污染。

我国运筹学的应用是在1957年始于建筑业和纺织业。在理论联系实际的思想指导下，从1958年开始在交通运输、工业、农业、水利建设、邮电等方面都有应用。尤其是在运输方面，从物资调运、装卸到调度等。在粮食部门为解决合理粮食调运问题，提出了“图上作业法”。我国的运筹学工作者从理论上证明了它的科学性。在解决邮递员合理投递路线时，管梅谷提出了国外称之为“中国邮路问题”的解法。在工业生产中推广了合理下料，机床负荷分配。在纺织业中曾用排队论方法解决细纱车间劳动组织，最优折布长度等问题。在农业中研究了作业布局、劳力分配和麦场设置等。从20世纪60年代起我国的运筹学工作者在钢铁和石油部门开展较全面的和深入的应用；投入产出法在钢铁部门首先得到应用。从1965年起统筹法的应用在建筑业、大型设备维修计划等方面取得可喜的进展。从1970年起在全国大部分省、市和部门推广优选法。其应用范围有配方、配比的选

择、生产工艺条件的选择、工艺参数的确定、工程设计参数的选择、仪器仪表的调试等。在20世纪70年代中期最优化方法在工程设计界得到广泛的重视。在光学设计、船舶设计、飞机设计、变压器设计、电子线路设计、建筑结构设计和化工过程设计等方面都有成果。从20世纪70年代中期排队论开始应用于研究矿山、港口、电讯和计算机的设计等方面。图论曾用于线路布置和计算机的设计、化学物品的存放等。存储论在我国应用较晚,20世纪70年代末在汽车工业和其他部门取得成功。近年来运筹学的应用已趋向研究规模大和复杂的问题,如部门计划、区域经济规划等;并已与系统工程难以分解。

国际运筹学联合会为了表彰发展中国家运筹学应用好的项目设立了运筹进展奖,下面列举近年来中国曾获该奖的项目。

(1) 中国国家经济信息系统的项目评价系统(章祥荪,崔晋川,中科院应用数学所,1996)。

(2) 长江上游生态、经济、发展的最优化(刘光中、徐玖平等,成都科技大学,1996)。

(3) 中国粮食产量预测研究(陈锡康,潘晓明,杨翠红,中科院系统科学所,1999)。

(4) 运筹学在中国农业管理中应用(赵庆祯,山东师范大学,曲阜师范大学,1999)。

(5) 北京公共交通区域运营组织模式的公共汽车运营总调度管理(沈吟东等,武汉科技学院,2005)。

中国运筹学会2008年颁发了中国运筹学会首届科学技术奖,越民义(中科院应用数学所)荣获首届中国运筹学科学技术奖,第二届2010年祁力群(中国香港理工大学)获该奖项。

1.6 运筹学的展望

关于运筹学将往哪个方向发展,从20世纪70年代起西方运筹学工作者有种种观点,至今还未说清。这里提出某些运筹学界的观点,供研究参考。美国前运筹学会主席邦特(S. Bonder)认为,运筹学应在三个领域发展:运筹学应用、运筹科学和运筹数学。并强调发展前两者,从整体讲应协调发展。事实上运筹数学到20世纪70年代已形成一系列强有力的分支,数学描述相当完善,这是一件好事。正是这一点使不少运筹学界的前辈认为,有些专家钻进运筹数学的深处,而忘掉了运筹学的原有特色,忽略了多学科的横向交叉联系和解决实际问题的客观需求。近几年来出现一种新的批评,指出有些人只迷恋于数学模型的精巧、复杂化,使用高深的数学工具,而不善于处理面临大量新的不易解决的实际问题。现代运筹学工作者面临的大量新问题是经济、技术、社会、生态和政治等因素交叉在一起的复杂系统。因此,从20世纪70年代末至80年代初不少运筹学家提出:要大家注意研究大系统,注意与系统分析相结合。美国科学院国际开发署写了一本书,其书名就把系统分析和运筹学并列。有的运筹学家提出了“要从运筹学到系统分析”的报告。由于研究新问题的时间跨度很长,因此必须与未来学紧密结合。由于面临的问题大多是

涉及技术、经济、社会、心理等综合因素的研究，在运筹学中除常用的数学方法以外，还引入一些非数学的方法和理论。曾在20世纪50年代写过“运筹学的数学方法”的美国运筹学家沙旦(T. L. Saaty)，他在20世纪70年代末提出了层次分析法(AHP)，并认为过去过分强调细巧的数学模型，可是它很难解决那些非结构性的复杂问题。因此宁可使用看起来是简单和粗糙的方法，加上决策者的正确判断，却能解决实际问题。切克兰特(P. B. Checkland)把传统的运筹学方法称为硬系统思考，它适用于解决那种结构明确的系统以及战术和技术性强的问题，而对于结构不明确的，有人参与活动的系统就不太胜任了。这就应采用软系统思考方法，相应的一些概念和方法都应有所变化，如将过分理想化的“最优解”换成“满意解”。过去把求得的“解”看做精确的、不能变的凝固的东西，而现在要以“易变性”的理念看待所得的“解”，以适应系统的不断变化。解决问题的过程是决策者和分析者发挥其创造性的过程，这就是进入20世纪70年代以来人们愈来愈对人机对话的算法感兴趣的原因。在20世纪80年代中一些重要的与运筹学有关的国际会议中，大多数认为决策支持系统是使运筹学发展的一个好机会。进入20世纪90年代和21世纪初期，发生两个很重要的趋势。一个是软运筹学崛起。主要发源地是在英国。1989年英国运筹学学会开了一个会议，会后由罗森汉特(J. Rosenhead)主编了一本论文集，后来被称为软运筹学的“圣经”。里面提到了不少新的属于软运筹的方法。如软系统方法论(SSM：Checkland)、战略假设表面化与检验(SAST：Mason & Mitroff)、战略选择(SC：Friend)、问题结构法(PSM：Bryant & Rosenhead)、超对策(hypergame：Benett)、亚对策(Metagame：Howard)、战略选择发展与分析(SODA：Eden)、生存系统模型(VSM：Beer)、对话式计划(IP：Ackoff)、批判式系统启发(CSH：Ulrich)等。2001年该书推出修订版，增加了很多实例。另一个趋势是与优化有关的，即软计算。这种方法不追求严格最优，具有启发式思路，并借用来自生物学、物理学和其他学科的思想来解寻优方法。其中最著名的有遗传算法(GA：Holland)、模拟退火(SA：Metropolis)、神经网络(NN)、模糊逻辑(FL：Zadeh)、进化计算(EC)、禁忌算法(TS)、蚁群优化(ACO：Dorigo)等。目前国际上已有世界软计算协会。2004年召开第9届国际会议。但都是在网络上开会，并且出版杂志：Applied soft computing。此外在一些老的分支方面，如线性规划也出现新的亮点，如内点法；图论中出现无标度网络(scale-free network)等。总之运筹学还在不断发展中，新的思想、观点和方法不断涌现。本书作为一本教材，所提供的一些运筹学思想和方法都是基本的，作为学习运筹学的读者必须掌握的知识。

参考资料

[1] Moder J J, Elmaghraby S E. Handbook of Operations Research. Foundations and Fundamentals, Vol. 1; Models and Application, Vol. 2. Von Nostrand Reinhold Company, 1978

[2] Morse P M, Brown A A etc. Systems Analysis and Operations Research tool for policy and program

planning for developing countries,National Academy of Sciences,1976

[3] 日本运筹学会. OR 事典(1975); OR 事典(增补别册)OR 事例集(1983). 日科技连出版社

[4] Haley K B. Operational Research'78. North Holland Publish Company,1979

[5] Brans J P. Operational Research'81. North Holland Publish Company,1981

[6] Saaty T L. Mathematical Methods of Operations Research,McGraw Hill Book Company Inc. 1959

[7] Brans J P. Operational Research'84. North Holland Publish Company,1984

[8] Checkland P B. Systems thinking,systems practice. Wiley,Chichester,1981

[9] Rosenhead J. Mingers J. Rational analysis for a problematic world revisted. Problem structuring methods for complexity,uncertainty and conflict. Wiley,Chichester,2001

[10] 顾基发,唐锡晋. 软系统工程方法论与软运筹学. 系统研究. 杭州:浙江人民出版社,1996

Part 2 第 2 篇

线性规划与目标规划

线性规划是运筹学的一个重要分支。自 1947 年丹捷格(G. B. Dantzig)提出了一般线性规划问题求解的方法——单纯形法之后，线性规划在理论上趋向成熟，在实用中日益广泛与深入。特别是在电子计算机能处理成千上万个约束条件和决策变量的线性规划问题之后，线性规划的适用领域更为广泛了。从解决技术问题的最优化设计到工业、农业、商业、交通运输业、军事、经济计划和管理决策等领域线性规划都可以发挥作用，它已成了现代科学管理的重要手段之一。查恩斯(A. Charnes)与库伯(W. W. Cooper)继丹捷格之后，于 1961 年提出了目标规划，艾吉利(Y. Ijiri)提出了用优先因子来处理多目标问题，使目标规划得到发展。30 多年前，斯・姆・李(S. M. Lee)与杰斯开莱尼(V. Jaaskelainen)应用计算机处理目标规划问题，使目标规划在实际应用方面比线性规划更广泛，更为管理者所重视。

CHAPTER 2 第2章

线性规划与单纯形法

2.1 线性规划问题及其数学模型

2.1.1 问题的提出

在生产管理和经营活动中经常提出一类问题，即如何合理地利用有限的人力、物力、财力等资源，以便得到最好的经济效果。

例 1 某工厂在计划期内要安排生产Ⅰ、Ⅱ两种产品，已知生产单位产品所需的设备台时及 A、B 两种原材料的消耗，如表 2-1 所示。

表 2-1

资源＼产品	产品Ⅰ	产品Ⅱ	现有条件
设　备	1 台时/件	2 台时/件	8 台时
原材料 A	4kg/件	0	16kg
原材料 B	0	4kg/件	12kg

该工厂每生产一件产品Ⅰ可获利 2 元，每生产一件产品Ⅱ可获利 3 元，问应如何安排计划使该工厂获利最多？这问题可以用以下的数学模型来描述，设 x_1、x_2 分别表示在计划期内产品Ⅰ、Ⅱ的产量。因为设备的有效台时是 8，这是一个限制产量的条件，所以在确定产品Ⅰ、Ⅱ的产量时，要考虑不超过设备的有效台时数，即可用不等式表示为：

$$x_1+2x_2\leqslant 8$$

同理，因原材料 A、B 的限量，可以得到以下不等式

$$4x_1 \leqslant 16$$

$$4x_2\leqslant 12$$

该工厂的目标是在不超过所有资源限量的条件下，如何确定产量 x_1、x_2 以得到最大的利润。若用 z 表示利润，这时 $z=2x_1+3x_2$。综合上述，该计划问题可用数学模型表示为

目标函数　　$\max\ z=2x_1+3x_2$

满足约束条件　　$\begin{cases} x_1+2x_2\leqslant 8 \\ 4x_1 \leqslant 16 \\ 4x_2\leqslant 12 \\ x_1,x_2\geqslant 0 \end{cases}$

例 2 靠近某河流有两个化工厂(见图 2-1),流经第一化工厂的河流流量为每天 500 万立方米,在两个工厂之间有一条流量为每天 200 万立方米的支流。第一化工厂每天排放含有某种有害物质的工业污水 2 万立方米,第二化工厂每天排放这种工业污水 1.4 万立方米。从第一化工厂排出的工业污水流到第二化工厂以前,有 20%可自然净化。根据环保要求,河流中工业污水的含量应不大于 0.2%。这两个工厂都需各自处理一部分工业污水。第一化工厂处理工业污水的成本是 1 000 元/万立方米,第二化工厂处理工业污水的成本是 800 元/万立方米。现在要问在满足环保要求的条件下,每厂各应处理多少工业污水,使这两个工厂总的处理工业污水费用最小。

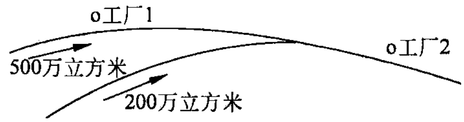


图 2-1

这个问题可用数学模型来描述。设第一化工厂每天处理工业污水量为 x_1 万立方米,第二化工厂每天处理工业污水量为 x_2 万立方米,从第一化工厂到第二化工厂之间,河流中工业污水含量要不大于 0.2%,由此可得近似关系式 $(2-x_1)/500\leqslant 2/1\,000$。

流经第二化工厂后,河流中的工业污水量仍要不大于 0.2%,这时有近似关系式

$$[0.8(2-x_1)+(1.4-x_2)]/700\leqslant 2/1\,000$$

由于每个工厂每天处理的工业污水量不会大于每天的排放量,故有 $x_1\leqslant 2$; $x_2\leqslant 1.4$。这问题的目标是要求两厂用于处理工业污水的总费用最小,即 $z=1\,000x_1+800x_2$。综合上述,这个环保问题可用数学模型表示为

$$\text{目标函数}\quad \min z=1\,000x_1+800x_2$$

$$\text{满足约束条件}\quad \begin{cases} x_1 \geqslant 1 \\ 0.8x_1+x_2\geqslant 1.6 \\ x_1 \leqslant 2 \\ x_2\leqslant 1.4 \\ x_1,x_2 \geqslant 0 \end{cases}$$

从以上两例可以看出,它们都是属于一类优化问题。它们的共同特征:

(1) 每一个问题都用一组决策变量 $(x_1,x_2,\cdots,x_n)$ 表示某一方案,这组决策变量的值就代表一个具体方案。一般这些变量取值是非负且连续的。

(2) 要有建模的有关数据,如资源拥有量、消耗资源定额、创造新价值量等,并构成互不矛盾的约束条件,这些约束条件可以用一组线性等式或线性不等式来表示。

(3) 都有一个要求达到的目标,它可用决策变量及其有关的价值系数构成的线性函数(称为目标函数)来表示。按问题的不同,要求目标函数实现最大化或最小化。

满足以上三个条件的数学模型称为线性规划的数学模型。其一般形式为

$$\text{目标函数}\ \max(\min)\ z=c_1x_1+c_2x_2+\cdots+c_nx_n \tag{2-1}$$

$$\text{满足约束条件}\begin{cases}\left.\begin{array}{l}a_{11}x_1+a_{12}x_2+\cdots+a_{1n}x_n\leqslant(=,\geqslant)b_1\\a_{21}x_1+a_{22}x_2+\cdots+a_{2n}x_n\leqslant(=,\geqslant)b_2\\\cdots\\a_{m1}x_1+a_{m2}x_2+\cdots+a_{mn}x_n\leqslant(=,\geqslant)b_m\end{array}\right\} & (2\text{-}2)\\ x_1,\quad x_2,\ \cdots,\quad x_n\geqslant 0 & (2\text{-}3)\end{cases}$$

在线性规划的数学模型中，式(2-1)称为目标函数；c_j 为价值系数；式(2-2)、式(2-3)称为约束条件；a_{ij}称为技术系数，b_i 称为限额系数；式(2-3)也称为变量的非负约束条件。

一般情况下 $m<n$。线性规划问题模型是建立在以下隐含的重要假设基础上的。

(1) 比例性。是指每个决策变量 $x_j(j=1,2,\cdots,n)$在约束条件中与在目标函数中数值变化时，按 x_j对应的技术系数a_{ij}与价值系数c_j严格地成比例变化。如例1，每生产1件产品Ⅰ，可获利2元；生产10件，获利20元；生产100件，获利200元。这里不存在在实际经济活动中的边际效用递减效应。

(2) 可加性。即目标函数的总值是各组成部分值(c_jx_j)之和；第 i 个约束关系式中各组成部分值($a_{ij}x_j$)之和就是第 i 项资源需求总值。决策变量是独立的，决策变量之间不发生关联，且不允许变量之间有交叉。如规划广告预期收益时不能选择看球赛的观众为决策变量 x_1，看电影的观众为决策变量 x_2等。

(3) 可分性。决策变量的值具有可分性，即允许非整数值。

(4) 确定性。即 c_j，a_{ij}，b_i都是确定的已知值。

可见线性规划模型是实际经济管理问题的抽象与近似。

2.1.2 图解法

图解法简单直观，有助于了解线性规划问题求解的基本原理。现对上述例1用图解法求解。在以 x_1、x_2 为坐标轴的直角坐标系中，非负条件 $x_1,x_2\geqslant0$ 是指第一象限。例1的每个约束条件都代表一个半平面。如约束条件 $x_1+2x_2\leqslant8$ 是代表以直线 $x_1+2x_2=8$ 为边界的左下方的半平面，若同时满足 $x_1,x_2\geqslant0$，$x_1+2x_2\leqslant8$，$4x_1\leqslant16$ 和 $4x_2\leqslant12$ 的约束条件的点，必然落在 x_1，x_2 坐标轴和由这三个半平面交成的区域内。由例1的所有约束条件为半平面交成的区域见图2-2中的阴影部分。阴影区域中的每一个点(包括边界点)都是这个线性规划问题的解(称可行解)，因而此区域是例1的线性规划问题的解集合，称它为可行域。

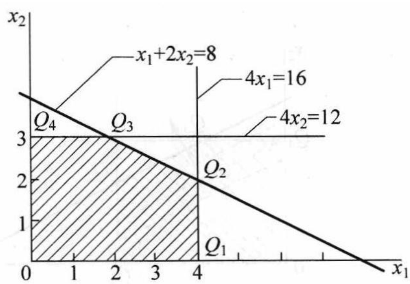


图 2-2

再分析目标函数 $z=2x_1+3x_2$，在这坐标平面上，它可表示以 z 为参数、$-2/3$ 为斜率的一族平行线

$$x_2=-(2/3)x_1+z/3$$

位于同一直线上的点，具有相同的目标函数值，因而称它为"等值线"。当 z 值由小变大时，直线 $x_2=-(2/3)x_1+z/3$ 沿其法线方向向右上方移动(见图 2-3 中箭线方向)。当移动到 Q_2 点时，使 z 值在可行域边界上实现最大化(见图 2-3)，这就得到了例 1 的最优解 Q_2，Q_2 点的坐标为(4,2)。于是可计算出满足所有约束条件下的最大值 $z=14$。

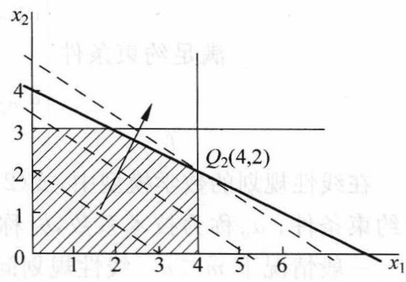


图 2-3

这说明该厂的最优生产计划方案是：生产 4 件产品Ⅰ，生产 2 件产品Ⅱ，可得最大利润为 14 元。

上例中求解得到问题的最优解是唯一的，但对一般线性规划问题，求解结果还可能出现以下几种情况：

1. 无穷多最优解(多重最优解)

若将例 1 中的目标函数变为求 $\max z=2x_1+4x_2$，则表示目标函数中以参数 z 的这族平行直线与约束条件 $x_1+2x_2\leqslant 8$ 的边界线平行。当 z 值由小变大时，将与线段 Q_2Q_3 重合(见图 2-4)。线段 Q_2Q_3 上任意一点都使 z 取得相同的最大值，这个线性规划问题有无穷多最优解(多重最优解)。

2. 无界解

对下述线性规划问题

$$\max z=x_1+x_2$$
$$\begin{cases}-2x_1+x_2\leqslant 4\\ x_1-x_2\leqslant 2\\ x_1,\ x_2\geqslant 0\end{cases}$$

用图解法求解结果见图 2-5。从图 2-5 中可以看到，该问题可行域无界，目标函数值可以增大到无穷大。称这种情况为无界解。

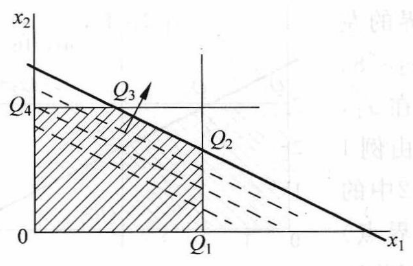


图 2-4

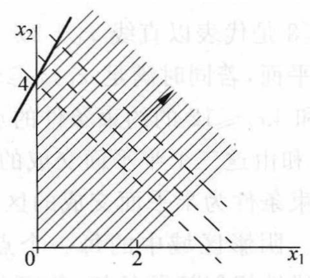


图 2-5

3. 无可行解

如果在例 1 的数学模型中增加一个约束条件 $-2x_1+x_2\geqslant 4$，该问题的可行域为空集，

既无可行解，也不存在最优解。

当求解结果出现第 2、3 两种情况时，一般说明线性规划问题的数学模型有错误。前者缺乏必要的约束条件，后者是有矛盾的约束条件，建模时应注意。

从图解法中直观地见到，当线性规划问题的可行域非空时，它是有界或无界凸多边形。若线性规划问题存在最优解，它一定在可行域的某个顶点得到；若在两个顶点同时得到最优解，则它们连线上的任意一点都是最优解，即有无穷多最优解。

图解法虽然直观、简便，但当变量数多于三个以上时，它就无能为力了。所以在第 3 节中要介绍一种代数法——单纯形法。为了便于讨论，先规定线性规划问题的数学模型的标准形式。

2.1.3 线性规划问题的标准形式

由前节可知，线性规划问题有各种不同的形式。目标函数有的要求 max，有的要求 min；约束条件可以是"≤"，也可以是"≥"形式的不等式，还可以是等式。决策变量一般是非负约束，但也允许取负值或在$(-\infty,\infty)$范围内取值，即无约束。将这些多种形式的数学模型统一变换为标准形式。这里规定的标准形式为(以下省略目标函数、约束条件的标记)。

$$(M_1)\qquad \max z=c_1x_1+c_2x_2+\cdots+c_nx_n$$

$$\begin{cases} a_{11}x_1+a_{12}x_2+\cdots+a_{1n}x_n=b_1 \\ a_{21}x_1+a_{22}x_2+\cdots+a_{2n}x_n=b_2 \\ \qquad\cdots \\ a_{m1}x_1+a_{m2}x_2+\cdots+a_{mn}x_n=b_m \\ x_1,x_2,\cdots,x_n\geqslant 0 \end{cases}$$

$$(M_1')\qquad \max z=\sum_{j=1}^{n}c_jx_j$$

$$\begin{cases} \sum_{j=1}^{n}a_{ij}x_j=b_i, & i=1,2,\cdots,m \\ x_j\geqslant 0, & j=1,2,\cdots,n \end{cases}$$

在标准形式中规定各约束条件的右端项 $b_i>0$，否则等式两端乘以"-1"。当某一个 $b_i=0$ 时，表示出现退化，这点将在以后讨论。

用向量和矩阵符号表述时为

$$(M_1'')\qquad \max z=\boldsymbol{CX}$$

$$\begin{cases} \sum_{j=1}^{n}\boldsymbol{P}_jx_j=\boldsymbol{b} \\ x_j\geqslant 0, \quad j=1,2,\cdots,n \end{cases}$$

并且约定：$\boldsymbol{C}=(c_1,c_2,\cdots,c_n)$，向量 $\boldsymbol{P}_j$ 对应的决策变量是 x_j。

$$\boldsymbol{X}=\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix};\quad \boldsymbol{P}_j=\begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{bmatrix};\quad \boldsymbol{b}=\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

用矩阵描述时为

$$\max z=\boldsymbol{CX}$$
$$\boldsymbol{AX}=\boldsymbol{b}$$
$$\boldsymbol{X}\geqslant 0$$

其中

$$\boldsymbol{A}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}=(\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_n);\quad 0=\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$\boldsymbol{A}$——约束条件的 $m\times n$ 维系数矩阵，一般有 $m<n$；

$\boldsymbol{b}$——资源向量；

$\boldsymbol{C}$——价值向量；

$\boldsymbol{X}$——决策变量向量。

应用 2.3 节单纯形法求解各种线性规划问题的数学模型都应先变换为标准型式。

以下讨论如何化为标准型的问题。

(1) 若要求目标函数实现最小化，即 $\min z=\boldsymbol{CX}$。这时只需将目标函数最小化变换求目标函数最大化，即令 $z'=-z$，于是得到 $\max z'=-\boldsymbol{CX}$。这就同标准型的目标函数的形式一致了。

(2) 约束方程为不等式。这里有两种情况：一种是约束方程为“$\leqslant$”不等式，则可在“$\leqslant$”不等式的左端加入非负松弛变量，把原“$\leqslant$”不等式变为等式；另一种是约束方程为“$\geqslant$”不等式，则可在“$\geqslant$”不等式的左端减去一个非负剩余变量(也可称松弛变量)，把不等式约束条件变为等式约束条件。下面举例说明。

例 3 将例 1 的数学模型化为标准型。

例 1 的数学模型(简称模型 M_2)为

$$\max z=2x_1+3x_2$$

$$\begin{cases} x_1+2x_2\leqslant 8 \\ 4x_1 \qquad\ \leqslant 16 \\ \qquad 4x_2\leqslant 12 \\ x_1,\ x_2\geqslant 0 \end{cases}$$

在各不等式中分别加上一个松弛变量 x_3,x_4,x_5，使不等式变为等式。这时得到标准型

$$\max z=2x_1+3x_2+0x_3+0x_4+0x_5$$

$$\begin{cases} x_1+2x_2+x_3 & =8 \\ 4x_1 \qquad +x_4 & =16 \\ \qquad 4x_2 \qquad +x_5 & =12 \\ x_1,\ x_2,\ x_3,x_4,x_5\geqslant 0 \end{cases}$$

所加松弛变量 x_3、x_4、x_5 表示没有被利用的资源。当然也没有利润，在目标函数中其系数应为零，即 $c_3,c_4,c_5=0$。

(3) 若存在取值无约束的变量 x_k，可令 $x_k=x'_k-x''_k$，其中 $x'_k,x''_k\geqslant 0$。

以上讨论说明，任何形式的数学模型都可化为标准型，下面举例说明。

例 4 将下述线性规划问题化为标准型

$$\min z=-x_1+2x_2-3x_3$$

$$\begin{cases} x_1+x_2+x_3 & \leqslant 7 \\ x_1-x_2+x_3 & \geqslant 2 \\ -3x_1+x_2+2x_3 & =5 \\ x_1,x_2\geqslant 0, & x_3\text{ 为无约束} \end{cases}$$

解 步骤：

(1) 用 x_4-x_5 替换 x_3，其中 $x_4,x_5\geqslant 0$；

(2) 在第一个约束不等式≤号的左端加入松弛变量 x_6；

(3) 在第二个约束不等式≥号的左端减去剩余变量 x_7；

(4) 令 $z'=-z$，把求 $\min z$ 改为求 $\max z'$，即可得到该问题的标准型

$$\max z'=x_1-2x_2+3(x_4-x_5)+0x_6+0x_7$$

$$\begin{cases} x_1+x_2+(x_4-x_5)+x_6 & =7 \\ x_1-x_2+(x_4-x_5)-x_7 & =2 \\ -3x_1+x_2+2(x_4-x_5) & =5 \\ x_1,x_2,x_4,x_5,x_6,x_7\geqslant 0 \end{cases}$$

2.1.4 线性规划问题解的概念

在讨论线性规划问题的求解前，先要了解线性规划问题的解的概念。由 2.1.3 节的 (M_1) 可知，一般线性规划问题的标准型为

$$\max z=\sum_{j=1}^{n}c_jx_j \tag{2-4}$$

$$\begin{cases} \sum_{j=1}^{n}a_{ij}x_j=b_i, & i=1,2,\cdots,m \quad (2\text{-}5) \\ x_j\geqslant 0, & j=1,2,\cdots,n \quad (2\text{-}6) \end{cases}$$

1. 可行解

满足约束条件式(2-5)、式(2-6)的解 $\boldsymbol{X}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$，称为线性规划问题的可行解，其中使目标函数达到最大值的可行解称为最优解。

2. 基

设 A 是约束方程组的 $m\times n(m<n)$ 维系数矩阵，其秩为 m。$\boldsymbol{B}$ 是矩阵 $\boldsymbol{A}$ 中 $m\times m$ 阶非奇异子矩阵($|\boldsymbol{B}|\neq 0$)，则称 $\boldsymbol{B}$ 是线性规划问题的一个基。这就是说，矩阵 $\boldsymbol{B}$ 是由 m 个线性独立的列向量组成。为不失一般性，可设

$$\boldsymbol{B}=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{pmatrix}=(\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_m)$$

称 $\boldsymbol{P}_j(j=1,2,\cdots,m)$ 为基向量，与基向量 $\boldsymbol{P}_j$ 相应的变量 $x_j(j=1,2,\cdots,m)$ 为基变量，否则称为非基变量，为了进一步讨论线性规划问题的解，下面研究约束方程组(2-5)的求解问题。假设该方程组系数矩阵 $\boldsymbol{A}$ 的秩为 m，因 $m<n$，故它有无穷多个解。假设前 m 个变量的系数列向量是线性独立的。这时式(2-5)可写成

$$\begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{pmatrix}x_1+\begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{pmatrix}x_2+\cdots+\begin{pmatrix} a_{1m} \\ a_{2m} \\ \vdots \\ a_{mm} \end{pmatrix}x_m$$

$$=\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}-\begin{pmatrix} a_{1,m+1} \\ a_{2,m+1} \\ \vdots \\ a_{m,m+1} \end{pmatrix}x_{m+1}-\cdots-\begin{pmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{pmatrix}x_n \tag{2-7}$$

或

$$\sum_{j=1}^{m}\boldsymbol{P}_j x_j=\boldsymbol{b}-\sum_{j=m+1}^{n}\boldsymbol{P}_j x_j$$

方程组(2-7)的一个基是

$$\boldsymbol{B}=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{pmatrix}=(\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_m)$$

设 $\boldsymbol{X}_B$ 是对应于这个基的基变量

$$\boldsymbol{X}_B=(x_1,x_2,\cdots,x_m)^{\mathrm{T}}$$

现若令式(2-7)的非基变量 $x_{m+1}=x_{m+2}=\cdots=x_n=0$，这时变量的个数等于线性方程的个数。用高斯消去法，求出一个解

$$\boldsymbol{X}=(x_1,x_2,\cdots,x_m,0,\cdots,0)^{\mathrm{T}}$$

该解的非零分量的数目不大于方程个数 m，称 $\boldsymbol{X}$ 为基解。由此可见，有一个基，就可以求出一个基解。如图 2-2 中的点 0，Q_1，Q_2，Q_3，Q_4 以及延长各条线（包括 $x_1=0$，$x_2=0$）的交点都代表基解。

3. 基可行解

满足非负条件(2-6)的基解，称为基可行解。图 2-2 中的点 0，Q_1，Q_2，Q_3，Q_4 代表基可行解。可见，基可行解的非零分量的数目也不大于 m，并且都是非负的。

4. 可行基

对应于基可行解的基，称为可行基。约束方程组(2-5)具有基解的数目最多是 C_n^m 个。一般基可行解的数目要小于基解的数目。以上提到的几种解的概念，它们之间的关系可用图 2-6 表明。另外还要说明一点，基解中的非零分量的个数小于 m 个时，该基解是退化解。在以下讨论时，假设不出现退化的情况。以上给出了线性规划问题的解的概念和定义，它们将有助于用来分析线性规划问题的求解过程。

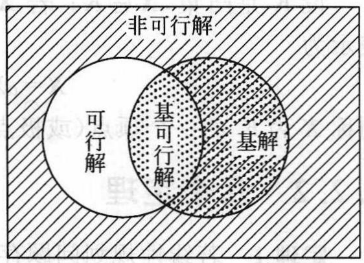


图 2-6

2.2 线性规划问题的几何意义

在 2.1.2 节介绍图解法时，已直观地看到可行域和最优解的几何意义，这一节从理论上进一步讨论。

2.2.1 基本概念

1. 凸集

设 K 是 n 维欧氏空间的一点集，若任意两点 $\boldsymbol{X}^{(1)}\in K$，$\boldsymbol{X}^{(2)}\in K$ 的连线上的所有点 $\alpha\boldsymbol{X}^{(1)}+(1-\alpha)\boldsymbol{X}^{(2)}\in K$，$(0\leqslant\alpha\leqslant1)$；则称 K 为凸集。

实心圆，实心球体，实心立方体等都是凸集，圆环不是凸集，从直观上讲，凸集没有凹入部分，其内部没有空洞。图 2-7 中的(a)(b)是凸集，(c)不是凸集。图 2-2 中的阴影部分是凸集。任何两个凸集的交集是凸集，见图 2-7(d)。

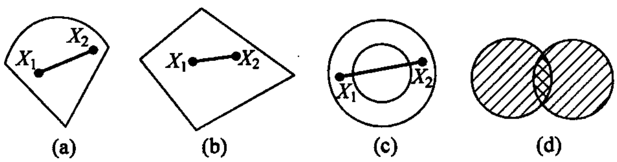


图 2-7

2. 凸组合

设 $\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)},\cdots,\boldsymbol{X}^{(k)}$ 是 n 维欧氏空间 E^n 中的 k 个点。若存在 $\mu_1,\mu_2,\cdots,\mu_k$，且 $0\leqslant \mu_i\leqslant 1,i=1,2,\cdots,k$；$\sum_{i=1}^{k}\mu_i=1$，使

$$\boldsymbol{X}=\mu_1\boldsymbol{X}^{(1)}+\mu_2\boldsymbol{X}^{(2)}+\cdots+\mu_k\boldsymbol{X}^{(k)}$$

则称 $\boldsymbol{X}$ 为 $\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)},\cdots,\boldsymbol{X}^{(k)}$ 的凸组合。(当 $0<\mu_i<1$ 时，称为严格凸组合)

3. 顶点

设 K 是凸集，$\boldsymbol{X}\in K$；若 $\boldsymbol{X}$ 不能用不同的两点 $\boldsymbol{X}^{(1)}\in K$ 和 $\boldsymbol{X}^{(2)}\in K$ 的线性组合表示为

$$\boldsymbol{X}=\alpha\boldsymbol{X}^{(1)}+(1-\alpha)\boldsymbol{X}^{(2)},\quad (0<\alpha<1)$$

则称 $\boldsymbol{X}$ 为 K 的一个顶点(或极点)。

2.2.2 几个定理

定理 1 若线性规划问题存在可行域，则其可行域

$$D=\left\{\boldsymbol{X}\,\middle|\,\sum_{j=1}^{n}\boldsymbol{P}_j x_j=\boldsymbol{b},\quad x_j\geqslant 0\right\}\text{是凸集}$$

证 为了证明满足线性规划问题的约束条件

$$\sum_{j=1}^{n}\boldsymbol{P}_j x_j=\boldsymbol{b},x_j\geqslant 0,\quad j=1,2,\cdots,n$$

的所有点(可行解)组成的集合是凸集，只要证明 D 中任意两点连线上的点必然在 D 内即可。

设
$$\boldsymbol{X}^{(1)}=(x_1^{(1)},x_2^{(1)},\cdots,x_n^{(1)})^{\mathrm{T}}$$
$$\boldsymbol{X}^{(2)}=(x_1^{(2)},x_2^{(2)},\cdots,x_n^{(2)})^{\mathrm{T}}$$

是 D 内的任意两点；$\boldsymbol{X}^{(1)}\neq\boldsymbol{X}^{(2)}$。

则有

$$\sum_{j=1}^{n}\boldsymbol{P}_j x_j^{(1)}=\boldsymbol{b},\quad x_j^{(1)}\geqslant 0,\quad j=1,2,\cdots,n$$

$$\sum_{j=1}^{n}\boldsymbol{P}_j x_j^{(2)}=\boldsymbol{b},\quad x_j^{(2)}\geqslant 0,\quad j=1,2,\cdots,n$$

令 $\boldsymbol{X}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$ 为 $x^{(1)},x^{(2)}$ 连线上的任意一点，即

$$\boldsymbol{X}=\alpha\boldsymbol{X}^{(1)}+(1-\alpha)\boldsymbol{X}^{(2)}\quad(0\leqslant\alpha\leqslant 1)$$

$\boldsymbol{X}$ 的每一个分量是 $x_j=\alpha x_j^{(1)}+(1-\alpha)x_j^{(2)}$，将它代入约束条件，得到

$$\begin{aligned}\sum_{j=1}^{n}\boldsymbol{P}_j x_j&=\sum_{j=1}^{n}\boldsymbol{P}_j[\alpha x_j^{(1)}+(1-\alpha)x_j^{(2)}]\\&=\alpha\sum_{j=1}^{n}\boldsymbol{P}_j x_j^{(1)}+\sum_{j=1}^{n}\boldsymbol{P}_j x_j^{(2)}-\alpha\sum_{j=1}^{n}\boldsymbol{P}_j x_j^{(2)}\end{aligned}$$

$$= \alpha \boldsymbol{b} + \boldsymbol{b} - \alpha \boldsymbol{b} = \boldsymbol{b}$$

又因 $x_j^{(1)}, x_j^{(2)} \geqslant 0, \alpha > 0, 1-\alpha > 0$，所以 $x_j \geqslant 0, j=1,2,\cdots,n$。由此可见 $\boldsymbol{X} \in D$，D 是凸集。

引理 1 线性规划问题的可行解 $\boldsymbol{X}=(x_1, x_2, \cdots, x_n)^{\mathrm{T}}$ 为基可行解的充要条件是 $\boldsymbol{X}$ 的正分量所对应的系数列向量是线性独立的。

证 (1) 必要性 由基可行解的定义可知。

(2) 充分性 若向量 $\boldsymbol{P}_1, \boldsymbol{P}_2, \cdots, \boldsymbol{P}_k$ 线性独立，则必有 $k \leqslant m$；当 $k=m$ 时，它们恰构成一个基，从而 $\boldsymbol{X}=(x_1, x_2, \cdots, x_k, 0\cdots 0)$ 为相应的基可行解。当 $k<m$ 时，则一定可以从其余的列向量中取出 $m-k$ 个与 $\boldsymbol{P}_1, \boldsymbol{P}_2, \cdots, \boldsymbol{P}_k$ 构成最大的线性独立向量组，其对应的解恰为 $\boldsymbol{X}$，所以根据定义它是基可行解。

定理 2 线性规划问题的基可行解 $\boldsymbol{X}$ 对应于可行域 D 的顶点。

证 不失一般性，假设基可行解 $\boldsymbol{X}$ 的前 m 个分量为正。故

$$\sum_{j=1}^{m} \boldsymbol{P}_j x_j = \boldsymbol{b} \tag{2-8}$$

现在分两步来讨论，分别用反证法。

(1) 若 $\boldsymbol{X}$ 不是基可行解，则它一定不是可行域 D 的顶点

根据引理 1，若 $\boldsymbol{X}$ 不是基可行解，则其正分量所对应的系数列向量 $\boldsymbol{P}_1, \boldsymbol{P}_2, \cdots, \boldsymbol{P}_m$ 线性相关，即存在一组不全为零的数 $\alpha_i, i=1,2,\cdots,m$ 使得

$$\alpha_1 \boldsymbol{P}_1 + \alpha_2 \boldsymbol{P}_2 + \cdots + \alpha_m \boldsymbol{P}_m = 0 \tag{2-9}$$

用一个 $\mu>0$ 的数乘式(2-9)再分别与式(2-8)相加和相减，这样得到

$$(x_1 - \mu\alpha_1)\boldsymbol{P}_1 + (x_2 - \mu\alpha_2)\boldsymbol{P}_2 + \cdots + (x_m - \mu\alpha_m)\boldsymbol{P}_m = \boldsymbol{b}$$

$$(x_1 + \mu\alpha_1)\boldsymbol{P}_1 + (x_2 + \mu\alpha_2)\boldsymbol{P}_2 + \cdots + (x_m + \mu\alpha_m)\boldsymbol{P}_m = \boldsymbol{b}$$

现取

$$\boldsymbol{X}^{(1)} = [(x_1 - \mu\alpha_1), (x_2 - \mu\alpha_2), \cdots, (x_m - \mu\alpha_m), 0, \cdots, 0]$$

$$\boldsymbol{X}^{(2)} = [(x_1 + \mu\alpha_1), (x_2 + \mu\alpha_2), \cdots, (x_m + \mu\alpha_m), 0, \cdots, 0]$$

由 $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}$ 可以得到 $\boldsymbol{X} = \frac{1}{2}\boldsymbol{X}^{(1)} + \frac{1}{2}\boldsymbol{X}^{(2)}$，即 $\boldsymbol{X}$ 是 $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}$ 连线的中点。

另一方面，当 μ 充分小时，可保证

$$x_i \pm \mu\alpha_i \geqslant 0, \quad i=1,2,\cdots,m$$

即 $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}$ 是可行解。这证明了 $\boldsymbol{X}$ 不是可行域 D 的顶点。

(2) 若 $\boldsymbol{X}$ 不是可行域 D 的顶点，则它一定不是基可行解

因为 $\boldsymbol{X}$ 不是可行域 D 的顶点，故在可行域 D 中可找到不同的两点

$$\boldsymbol{X}^{(1)} = (x_1^{(1)}, x_2^{(1)}, \cdots, x_n^{(1)})^{\mathrm{T}}$$

$$\boldsymbol{X}^{(2)} = (x_1^{(2)}, x_2^{(2)}, \cdots, x_n^{(2)})^{\mathrm{T}}$$

使

$$\boldsymbol{X} = \alpha \boldsymbol{X}^{(1)} + (1-\alpha)\boldsymbol{X}^{(2)} \quad 0<\alpha<1$$

设 $\boldsymbol{X}$ 是基可行解，对应向量组 $\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_m$ 线性独立。当 $j>m$ 时，有 $x_j=x_j^{(1)}=x_j^{(2)}=0$，由于 $\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)}$ 是可行域的两点。应满足

$$\sum_{j=1}^{m}\boldsymbol{P}_j x_j^{(1)}=\boldsymbol{b} \quad 与 \quad \sum_{j=1}^{m}\boldsymbol{P}_j x_j^{(2)}=\boldsymbol{b}$$

将这两式相减，即得

$$\sum_{j=1}^{m}\boldsymbol{P}_j(x_j^{(1)}-x_j^{(2)})=0$$

因 $\boldsymbol{X}^{(1)}\neq\boldsymbol{X}^{(2)}$，所以上式系数 $(x_j^{(1)}-x_j^{(2)})$ 不全为零，故向量组 $\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_m$ 线性相关，与假设矛盾。即 $\boldsymbol{X}$ 不是基可行解。

引理 2 若 K 是有界凸集，则任何一点 $\boldsymbol{X}\in K$ 可表示为 K 的顶点的凸组合。

本引理证明从略，用以下例子说明这引理。

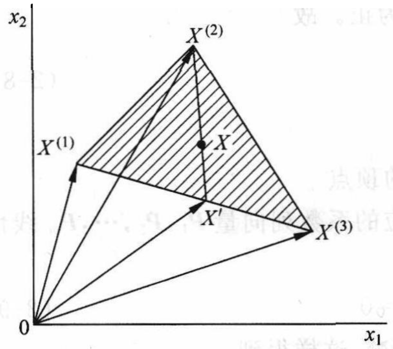


图 2-8

例 5 设 $\boldsymbol{X}$ 是三角形中任意一点，$\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)}$ 和 $\boldsymbol{X}^{(3)}$ 是三角形的三个顶点，试用三个顶点的坐标表示 $\boldsymbol{X}$（见图 2-8）。

解 任选一顶点 $\boldsymbol{X}^{(2)}$，做一条连线 $\boldsymbol{XX}^{(2)}$；并延长交于 $\boldsymbol{X}^{(1)}$、$\boldsymbol{X}^{(3)}$ 连接线上一点 $\boldsymbol{X}'$。因 $\boldsymbol{X}'$ 是 $\boldsymbol{X}^{(1)}$、$\boldsymbol{X}^{(3)}$ 连线上一点，故可用 $\boldsymbol{X}^{(1)}$、$\boldsymbol{X}^{(3)}$ 线性表示为

$$\boldsymbol{X}'=\alpha\boldsymbol{X}^{(1)}+(1-\alpha)\boldsymbol{X}^{(3)} \quad 0<\alpha<1$$

又因 $\boldsymbol{X}$ 是 $\boldsymbol{X}'$ 与 $\boldsymbol{X}^{(2)}$ 连线上的一个点，故

$$\boldsymbol{X}=\lambda\boldsymbol{X}'+(1-\lambda)\boldsymbol{X}^{(2)} \quad 0<\lambda<1$$

将 $\boldsymbol{X}'$ 的表达式代入上式得到

$$\begin{aligned}\boldsymbol{X}&=\lambda[\alpha\boldsymbol{X}^{(1)}+(1-\alpha)\boldsymbol{X}^{(3)}]+(1-\lambda)\boldsymbol{X}^{(2)}\\&=\lambda\alpha\boldsymbol{X}^{(1)}+\lambda(1-\alpha)\boldsymbol{X}^{(3)}+(1-\lambda)\boldsymbol{X}^{(2)}\end{aligned}$$

令 $\mu_1=\alpha\lambda,\mu_2=(1-\lambda),\mu_3=\lambda(1-\alpha)$

这就得到

$$\boldsymbol{X}=\mu_1\boldsymbol{X}^{(1)}+\mu_2\boldsymbol{X}^{(2)}+\mu_3\boldsymbol{X}^{(3)}$$

$$\sum_i\mu_i=1,\quad 0<\mu_i<1$$

定理 3 若可行域有界，线性规划问题的目标函数一定可以在其可行域的顶点上达到最优。

证 设 $\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)},\cdots,\boldsymbol{X}^{(k)}$ 是可行域的顶点，若 $\boldsymbol{X}^{(0)}$ 不是顶点，且目标函数在 $\boldsymbol{X}^{(0)}$ 处达到最优 $z^*=\boldsymbol{CX}^{(0)}$（标准型是 $z^*=\max z$）。

因 $\boldsymbol{X}^{(0)}$ 不是顶点，所以它可以用 D 的顶点线性表示为

$$\boldsymbol{X}^{(0)}=\sum_{i=1}^{k}\alpha_i\boldsymbol{X}^{(i)},\alpha_i>0,\sum_{i=1}^{k}\alpha_i=1$$

因此

$$CX^{(0)} = C\sum_{i=1}^{k}\alpha_i X^{(i)} = \sum_{i=1}^{k}\alpha_i CX^{(i)} \tag{2-10}$$

在所有的顶点中必然能找到某一个顶点 $X^{(m)}$，使 $CX^{(m)}$ 是所有 $CX^{(i)}$ 中最大者。并且将 $X^{(m)}$ 代替式(2-10)中的所有 $X^{(i)}$，这就得到

$$\sum_{i=1}^{k}\alpha_i CX^{(i)} \leqslant \sum_{i=1}^{k}\alpha_i CX^{(m)} = CX^{(m)}$$

由此得到

$$CX^{(0)} \leqslant CX^{(m)}$$

根据假设 $CX^{(0)}$ 是最大值，所以只能有

$$CX^{(0)} = CX^{(m)}$$

即目标函数在顶点 $X^{(m)}$ 处也达到最大值。

有时目标函数可能在多个顶点处达到最大值。这时在这些顶点的凸组合上也达到最大值。称这种线性规划问题有无穷多个最优解。

假设 $\hat{X}^{(1)},\hat{X}^{(2)},\cdots,\hat{X}^{(k)}$ 是目标函数达到最大值的顶点，若 $\hat{X}$ 是这些顶点的凸组合，即

$$\hat{X} = \sum_{i=1}^{k}\alpha_i \hat{X}^{(i)},\ \alpha_i > 0,\ \sum_{i=1}^{k}\alpha_i = 1$$

于是

$$C\hat{X} = C\sum_{i=1}^{k}\alpha_i \hat{X}^{(i)} = \sum_{i=1}^{k}\alpha_i C\hat{X}^{(i)}$$

设

$$C\hat{X}^{(i)} = m,\quad i=1,2,\cdots,k$$

于是

$$C\hat{X} = \sum_{i=1}^{k}\alpha_i m = m$$

另外，若可行域为无界，则可能无最优解，也可能有最优解，若有也必定在某顶点上得到。根据以上讨论，可以得到以下结论：

线性规划问题的所有可行解构成的集合是凸集，也可能为无界域，它们有有限个顶点，线性规划问题的每个基可行解对应可行域的一个顶点；若线性规划问题有最优解，必可在某顶点上得到。虽然顶点数目是有限的(它不大于 C_n^m 个)，若采用“枚举法”找所有基可行解，然后一一比较，最终可能找到最优解。但当 n,m 的数较大时，这种办法是行不通的，所以要继续讨论，如何有效地找到最优解，有多种方法，这里仅介绍单纯形法。

2.3 单纯形法

单纯形法求解线性规划的思路：一般线性规划问题具有线性方程组的变量数大于方程个数，这时有不定的解。单纯形法是在高斯消去法的基础上，发展为求解变量数多于方

程数，并且使目标函数值优化的方法。从线性方程组中找出一个个的单纯形，每一个单纯形可以求得一组解，然后再判断该解使目标函数值是增大还是变小，决定下一步选择的单纯形，这就是迭代。直到目标函数实现最大值或最小值为止，这样问题就得到了最优解。先举一例来说明。

注意：单纯形是指 0 维中的点，一维中的线段，二维中的三角形，三维中的四面体，n 维空间中的有 $n+1$ 个顶点的多面体。例如在三维空间中的四面体，其顶点分别为 $(0,0,0)$，$(1,0,0)$，$(0,1,0)$，$(0,0,1)$。具有单位截距的单纯形的方程是 $\sum x_i \leqslant 1$，并且 $x_i \geqslant 0, i=1,2,\cdots,m$。

2.3.1 举例

例 6 试以例 1 来讨论如何用单纯形法求解。例 1 的标准型为

$$\max z = 2x_1 + 3x_2 + 0x_3 + 0x_4 + 0x_5 \tag{2-11}$$

$$\begin{cases} x_1 + 2x_2 + x_3 = 8 \\ 4x_1 + x_4 = 16 \\ 4x_2 + x_5 = 12 \end{cases} \tag{2-12}$$

$$x_j \geqslant 0, \quad j=1,2,\cdots,5$$

约束方程(2-12)的系数矩阵

$$\boldsymbol{A} = (\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3, \boldsymbol{P}_4, \boldsymbol{P}_5) = \begin{pmatrix} 1 & 2 & 1 & 0 & 0 \\ 4 & 0 & 0 & 1 & 0 \\ 0 & 4 & 0 & 0 & 1 \end{pmatrix}$$

从式(2-12)中可以看到 x_3, x_4, x_5 的系数列向量

$$\boldsymbol{P}_3 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{P}_4 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \boldsymbol{P}_5 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

是线性独立的，这些向量构成一个基

$$\boldsymbol{B} = (\boldsymbol{P}_3, \boldsymbol{P}_4, \boldsymbol{P}_5) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

对应于 $\boldsymbol{B}$ 的变量 x_3, x_4, x_5 为基变量，从(2-12)式中可以得到

$$\begin{cases} x_3 = 8 - x_1 - 2x_2 \\ x_4 = 16 - 4x_1 \\ x_5 = 12 - 4x_2 \end{cases} \tag{2-13}$$

将式(2-13)代入目标函数式(2-11)得到

$$z = 0 + 2x_1 + 3x_2 \tag{2-14}$$

当令非基变量 $x_1=x_2=0$，便得到 $z=0$。这时得到一个基可行解 $\mathbf{X}^{(0)}$

$$\mathbf{X}^{(0)}=(0,0,8,16,12)^{\mathrm{T}}$$

这个基可行解表示：工厂没有安排生产产品Ⅰ、Ⅱ；资源都没有被利用，所以工厂的利润指标 $z=0$。

从分析目标函数的表达式(2-14)可以看到：非基变量 x_1,x_2（即没有安排生产产品Ⅰ，Ⅱ）的系数都是正数，因此将非基变量变换为基变量，目标函数的值就可能增大。从经济意义上讲，安排生产产品Ⅰ或Ⅱ，就可以使工厂的利润指标增加。所以只要在目标函数(2-14)的表达式中还存在有正系数的非基变量，这表示目标函数值还有增加的可能，就需要将非基变量与基变量进行对换。一般选择正系数最大的那个非基变量 x_2 为换入变量，将它换入到基变量中去，同时还要确定基变量中有一个要换出来成为非基变量，可按以下方法来确定换出变量。

现分析式(2-13)，当将 x_2 定为换入变量后，必须从 x_3,x_4,x_5 中确定一个换出变量，并保证其余的都是非负，即 $x_3,x_4,x_5\geqslant0$。

当 $x_1=0$，由式(2-13)得到

$$\begin{cases}x_3=\ 8-2x_2\geqslant0\\x_4=16\geqslant0\\x_5=12-4x_2\geqslant0\end{cases}\qquad(2\text{-}15)$$

从式(2-15)中可以看出，只有选择

$$x_2=\min(8/2,-,12/4)=3$$

时，才能使式(2-15)成立。因当 $x_2=3$ 时，基变量 $x_5=0$，这就决定用 x_2 去替换 x_5。以上数学描述说明了每生产一件产品Ⅱ，需要用掉各种资源数为(2,0,4)。由这些资源中的薄弱环节，就确定了产品Ⅱ的产量。这里就是由原材料 B 的数量确定了产品Ⅱ的产量

$$x_2=\frac{12}{4}=3(\text{件})$$

为了求得以 x_3,x_4,x_2 为基变量的一个基可行解和进一步分析问题，需将式(2-13)中 x_2 的位置与 x_5 的位置对换。得到

$$\begin{cases}x_3+2x_2=8-x_1 & ①\\x_4=16-4x_1 & ②\\4x_2=12-x_5 & ③\end{cases}\qquad(2\text{-}16)$$

用高斯消去法，将式(2-16)中 x_2 的系数列向量变换为单位列向量。其运算步骤是：③′=③/4；①′=①−2×③′；②′=②，并将结果仍按原顺序排列有：

$$\begin{cases}x_3=\ 2-\ x_1+\dfrac{1}{2}x_5 & ①'\\x_4=16-4x_1 & ②'\\x_2=\ 3\qquad-\dfrac{1}{4}x_5 & ③'\end{cases}\qquad(2\text{-}17)$$

再将式(2-17)代入目标函数式(2-11)得到

$$z=9+2x_1-\frac{3}{4}x_5 \tag{2-18}$$

令非基变量 $x_1=x_5=0$,得到 $z=9$,并得到另一个基可行解 $\boldsymbol{X}^{(1)}$

$$\boldsymbol{X}^{(1)}=(0,3,2,16,0)^{\mathrm{T}}$$

从目标函数的表达式(2-18)中可以看到,非基变量 x_1 的系数是正的,说明目标函数值还可以增大,$\boldsymbol{X}^{(1)}$ 不一定是最优解。于是再用上述方法,确定换入、换出变量,继续迭代,再得到另一个基可行解 $\boldsymbol{X}^{(2)}$

$$\boldsymbol{X}^{(2)}=(2,3,0,8,0)^{\mathrm{T}}$$

再经过一次迭代,再得到一个基可行解 $\boldsymbol{X}^{(3)}$

$$\boldsymbol{X}^{(3)}=(4,2,0,0,4)^{\mathrm{T}}$$

而这时得到目标函数的表达式

$$z=14-1.5x_3-0.125x_4 \tag{2-19}$$

再检查式(2-19),可见到所有非基变量 x_3,x_4 的系数都是负数。这说明若要用剩余资源 x_3,x_4,就必须支付附加费用。所以当 $x_3=x_4=0$ 时,即不再利用这些资源时,目标函数达到最大值。所以 $\boldsymbol{X}^{(3)}$ 是最优解。即当产品Ⅰ生产4件,产品Ⅱ生产2件,工厂才能得到最大利润。通过上例,可以了解利用单纯形法求解线性规划问题的思路。现将每步迭代得到的结果与图解法作一对比,其几何意义就很清楚了。

原例1的线性规划问题是二维的,即两个变量 x_1,x_2;当加入松弛变量 x_3,x_4,x_5 后,变换为高维的,这时可以想象,满足所有约束条件的可行域是高维空间的凸多面体(凸集)。这凸多面体上的顶点,就是基可行解。如回到平面坐标的情况,初始基可行解 $\boldsymbol{X}^{(0)}=(0,0,8,16,12)^{\mathrm{T}}$ 就相当于图2-2中的原点(0,0),$\boldsymbol{X}^{(1)}=(0,3,2,16,0)^{\mathrm{T}}$ 相当于图2-2中的 Q_4 点(0,3);$\boldsymbol{X}^{(2)}=(2,3,0,8,0)^{\mathrm{T}}$ 相当于图2-2中的 Q_3 点(2,3),最优解 $\boldsymbol{X}^{(3)}=(4,2,0,0,4)^{\mathrm{T}}$ 相当于图2-2中的 Q_2 点(4,2)。从初始基可行解 $\boldsymbol{X}^{(0)}$ 开始迭代,依次得到 $\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)},\boldsymbol{X}^{(3)}$。这相当于图2-2中的目标函数平移时,从0点开始,首先碰到 Q_4,然后碰到 Q_3,最后达到 Q_2。下面讨论一般线性规划问题的求解。

2.3.2 初始基可行解的确定

为了确定初始基可行解,要首先找出初始可行基,其方法如下。

1. 直观判断

若线性规划问题

$$\max z=\sum_{j=1}^{n}c_jx_j \tag{2-20}$$

$$\sum_{j=1}^{n}\boldsymbol{P}_jx_j=\boldsymbol{b} \tag{2-21}$$

$$x_j \geqslant 0, \quad j=1,2,\cdots,n$$

从 $\boldsymbol{P}_j(j=1,2,\cdots,n)$ 中一般能直接观察到存在一个初始可行基

$$\boldsymbol{B}=(\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_m)=\begin{bmatrix}1 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & 1\end{bmatrix}$$

2. 加松弛变量

对所有约束条件是“≤”形式的不等式，可以利用化为标准型的方法，在每个约束条件的左端加上一个松弛变量。经过整理，重新对 x_j 及 $a_{ij}(i=1,2,\cdots,m;\ j=1,2,\cdots,n)$ 进行编号，则可得下列方程组

$$\begin{cases} x_1 \qquad + a_{1,m+1}x_{m+1} + \cdots + a_{1n}x_n = b_1 \\ \quad x_2 \quad + a_{2,m+1}x_{m+1} + \cdots + a_{2n}x_n = b_2 \\ \qquad \cdots \\ \qquad x_m + a_{m,m+1}x_{m+1} + \cdots + a_{mn}x_n = b_m \\ x_j \geqslant 0, j = 1,2,\cdots,n \end{cases} \tag{2-22}$$

显然得到一个 $m\times m$ 单位矩阵

$$\boldsymbol{B}=(\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_m)=\begin{bmatrix}1 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & 1\end{bmatrix}$$

以 $\boldsymbol{B}$ 作为可行基。将式(2-22)每个等式移项得

$$\begin{cases} x_1 = b_1 - a_{1,m+1}x_{m+1} - \cdots - a_{1n}x_n \\ x_2 = b_2 - a_{2,m+1}x_{m+1} - \cdots - a_{2n}x_n \\ \qquad \cdots \\ x_m = b_m - a_{m,m+1}x_{m+1} - \cdots - a_{mn}x_n \end{cases} \tag{2-23}$$

令 $x_{m+1}=x_{m+2}=\cdots=x_n=0$，由式(2-23)可得

$$x_i=b_i \quad (i=1,2,\cdots,m)$$

又因 $b_i\geqslant 0$（在 2.1.3 节中已作过规定），所以得到一个初始基可行解

$$\begin{aligned}\boldsymbol{X} &= (x_1,x_2,\cdots,x_m,\underbrace{0,\cdots,0}_{n-m\text{个}})^{\mathrm{T}} \\ &= (b_1,b_2,\cdots,b_m,\underbrace{0,\cdots,0}_{n-m\text{个}})^{\mathrm{T}}\end{aligned}$$

3. 加人工变量

对所有约束条件是“≥”形式的不等式及等式约束情况，若不存在单位矩阵时，就采用人造基方法。即对不等式约束减去一个非负的剩余变量后，再加上一个非负的人工变量；

对于等式约束再加上一个非负的人工变量，总能得到一个单位矩阵。关于这个方法将在第 2.5 节中进一步讨论。

2.3.3 最优性检验与解的判别

对线性规划问题的求解结果可能出现唯一最优解、无穷多最优解、无界解和无可行解四种情况，为此需要建立对解的判别准则。一般情况下，经过迭代后式(2-23)变成

$$x_i = b'_i - \sum_{j=m+1}^{n} a'_{ij} x_j \quad (i = 1,2,\cdots,m) \tag{2-24}$$

将式(2-24)代入目标函数式(2-20)，整理后得

$$z = \sum_{i=1}^{m} c_i b'_i + \sum_{j=m+1}^{n} \left(c_j - \sum_{i=1}^{m} c_i a'_{ij}\right) x_j \tag{2-25}$$

令

$$z_0 = \sum_{i=1}^{m} c_i b'_i, z_j = \sum_{i=1}^{m} c_i a'_{ij}, \quad (j = m+1,\cdots,n)$$

于是

$$z = z_0 + \sum_{j=m+1}^{n} (c_j - z_j) x_j \tag{2-26}$$

再令

$$\sigma_j = c_j - z_j \quad (j = m+1,\cdots,n)$$

则

$$z = z_0 + \sum_{j=m+1}^{n} \sigma_j x_j \tag{2-27}$$

1. 最优解的判别规则

若 $\boldsymbol{X}^{(0)} = (b'_1, b'_2, \cdots, b'_m, 0, \cdots, 0)^{\mathrm{T}}$ 为对应于基 $\boldsymbol{B}$ 的一个基可行解，且对于所有 $j = m+1, \cdots, n$，有 $\sigma_j \leqslant 0$，则 $\boldsymbol{X}^{(0)}$ 为最优解。称 σ_j 为检验数。

当所有非基变量的 $\sigma_j \leqslant 0$ 时，由式(2-27)可知已不存在任一可换入的非基变量，使目标函数继续增大。所以以 $\sigma_j \leqslant 0$ 为最优解的判别准则。

2. 无穷多最优解判别规则

若 $\boldsymbol{X}^{(0)} = (b'_1, b'_2, \cdots, b'_m, 0, \cdots, 0)^{\mathrm{T}}$ 为一个基可行解，对于一切 $j = m+1, \cdots, n$，有 $\sigma_j \leqslant 0$，又存在某个非基变量的检验数 $\sigma_{m+k} = 0$，则线性规划问题有无穷多最优解。

证 只需将非基变量 x_{m+k} 换入基变量中，找到一个新基可行解 $\boldsymbol{X}^{(1)}$。因 $\sigma_{m+k} = 0$，由式(2-27)知 $z = z_0$，故 $\boldsymbol{X}^{(1)}$ 也是最优解。由 2.2.2 节的定理 3 可知 $\boldsymbol{X}^{(0)}, \boldsymbol{X}^{(1)}$ 连线上所有点都是最优解。

3. 无界解判别规则

若 $\boldsymbol{X}^{(0)} = (b'_1, b'_2, \cdots, b'_m, 0, \cdots, 0)^{\mathrm{T}}$ 为一基可行解，有一个 $\sigma_{m+k} > 0$，并且对 $i = 1$,

$2,\cdots,m$，有 $a'_{i,m+k}\leqslant 0$，那么该线性规划问题具有无界解(或称无最优解)。

因有一个 $\sigma_{m+k}>0$，它对应的非基变量 x_{m+k} 为换入变量，使目标函数值继续增大，但因 $a'_{i,m+k}\leqslant 0$，这时无换出变量。由式(2-24)可知，在约束条件 $x_j\geqslant 0$ 时，只有为 $x_j=0$ 时，$x_i=b_i$，否则 $x_i<b_i$；现在存在 $a'_{i,m+k}\leqslant 0$，x_j 取任何大于 0 的值，就有 $x_i>b$。

证 构造一个新的解 $\boldsymbol{X}^{(1)}$，它的分量为

$$x_i^{(1)}=b'_i-\lambda a'_{i,m+k}\quad(\lambda>0)$$

$$x_{m+k}^{(1)}=\lambda$$

$$x_j^{(1)}=0,\quad j=m+1,\cdots,n,\text{且 } j\neq m+k$$

因 $a'_{i,m+k}\leqslant 0$，所以对任意的 $\lambda>0$ 都是可行解，把 $x^{(1)}$ 代入目标函数内得

$$z=z_0+\lambda\sigma_{m+k}$$

因 $\sigma_{m+k}>0$，故当 $\lambda\to+\infty$，则 $z\to+\infty$，故该问题目标函数无界。

以上讨论都是针对标准型，即求目标函数极大化时的情况。当求目标函数极小化时，一种情况如前所述，将其化为标准型。如果不化为标准型，只需在上述 1，2 点中把 $\sigma_j\leqslant 0$ 改为 $\sigma_j\geqslant 0$，第 3 点中将 $\sigma_{m+k}>0$ 改写为 $\sigma_{m+k}<0$ 即可。

对求解结果为无可行解的判别将在 2.5.2 节中讨论。

2.3.4 基变换

若初始基可行解 $\boldsymbol{X}^{(0)}$ 不是最优解及不能判别无界时，需要找一个新的基可行解。具体做法是从原可行解基中换一个列向量(当然要保证线性独立)，得到一个新的可行基，这称为基变换。为了换基，先要确定换入变量，再确定换出变量，让它们相应的系数列向量进行对换，就得到一个新的基可行解。

1. 换入变量的确定

由式(2-27)看到，当某些 $\sigma_j>0$ 时，x_j 增加则目标函数值还可以增大，这时要将某个非基变量 x_j 换到基变量中去(称为换入变量)。若有两个以上的 $\sigma_j>0$，那么选哪个非基变量作为换入变量呢？为了使目标函数值增加得快，从直观上一般选 $\sigma_j>0$ 中的大者即

$$\max_j(\sigma_j>0)=\sigma_k$$

则对应的 x_k 为换入变量。但也可以任选或按最小足码选。

2. 换出变量的确定

设 $\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_m$ 是一组线性独立的向量组，它们对应的基可行解是 $\boldsymbol{X}^{(0)}$。将它代入约束方程组式(2-21)得到

$$\sum_{i=1}^{m}x_i^{(0)}\boldsymbol{P}_i=\boldsymbol{b}\tag{2-28}$$

其他的向量 $\boldsymbol{P}_{m+1},\boldsymbol{P}_{m+2},\cdots,\boldsymbol{P}_{m+t},\cdots,\boldsymbol{P}_n$ 都可以用 $\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_m$ 线性表示，若确定非基变量 $\boldsymbol{P}_{m+t}$ 为换入变量，必然可以找到一组不全为 0 的数$(i=1,2,\cdots,m)$使得

$$\boldsymbol{P}_{m+t}=\sum_{i=1}^{m}\beta_{i,m+t}\boldsymbol{P}_i$$

或

$$\boldsymbol{P}_{m+t}-\sum_{i=1}^{m}\beta_{i,m+t}\boldsymbol{P}_i=0 \tag{2-29}$$

在式(2-29)两边同乘一个正数 θ,然后将它加到式(2-28)上,得到

$$\sum_{i=1}^{m}x_i^{(0)}\boldsymbol{P}_i+\theta\left(\boldsymbol{P}_{m+t}-\sum_{i=1}^{m}\beta_{i,m+t}\boldsymbol{P}_i\right)=\boldsymbol{b}$$

或

$$\sum_{i=1}^{m}(x_i^{(0)}-\theta\beta_{i,m+t})\boldsymbol{P}_i+\theta\boldsymbol{P}_{m+t}=\boldsymbol{b} \tag{2-30}$$

当 θ 取适当值时,就能得到满足约束条件的一个新的基可行解(即非零分量的数目不大于 m 个)。就应使 $(x_i^{(0)}-\theta\beta_{i,m+t})(i=1,2,\cdots,m)$ 中的某一个为零,并保证其余的分量为非负。这个要求可以用以下的办法达到:比较各比值 $\dfrac{x_i^{(0)}}{\beta_{i,m+t}}(i=1,2,\cdots,m)$,因为 θ 必须是正数,所以只选择 $\left(\dfrac{x_i^{(0)}}{\beta_{i,m+t}}\right)>0(i=1,2,\cdots,m)$ 中比值最小的等于 θ。以上描述用数学式表示为

$$\theta=\min_i\left(\frac{x_i^{(0)}}{\beta_{i,m+t}}\middle|\beta_{i,m+t}>0\right)=\frac{x_l^{(0)}}{\beta_{l,m+t}}$$

这时 x_l 为换出变量。按最小比值确定 θ 值,称为最小比值规则或 θ 规划。将 $\theta=\dfrac{x_l^{(0)}}{\beta_{l,m+t}}$ 代入 $\boldsymbol{X}$ 中,便得到新的基可行解。

$$\boldsymbol{X}^{(1)}=\Big(x_1^{(0)}-\frac{x_l^{(0)}}{\beta_{l,m+t}}\cdot\beta_{1,m+t},\cdots,\underset{\substack{\uparrow\\ \text{第 } l \text{ 个分量}}}{0},\cdots,$$

$$x_m^{(0)}-\frac{x_l^{(0)}}{\beta_{l,m+t}}\cdot\beta_{m,m+t},0,\cdots,\underset{\substack{\uparrow\\ \text{第 } m+t \text{ 个分量}}}{\frac{x_l^{(0)}}{\beta_{l,m+t}}},\cdots,0\Big)$$

由此得到由 $\boldsymbol{X}^{(0)}$ 转换到 $\boldsymbol{X}^{(1)}$ 的各分量的转换公式

$$x_i^{(1)}=\begin{cases}x_i^{(0)}-\dfrac{x_l^{(0)}}{\beta_{l,m+t}}\cdot\beta_{i,m+t} & (i\neq l)\\[2ex] \dfrac{x_l^{(0)}}{\beta_{l,m+t}} & (i=l)\end{cases}$$

这里 $x_i^{(0)}$ 是原基可行解 $\boldsymbol{X}^{(0)}$ 的各分量;$x_i^{(1)}$ 是新基可行解 $\boldsymbol{X}^{(1)}$ 的各分量;$\beta_{i,m+t}$ 是换入向量 $\boldsymbol{P}_{m+t}$ 的对应原来一组基向量的坐标。现在的问题是,这个新解 $\boldsymbol{X}^{(1)}$ 的 m 个非零分量对应的列向量是否线性独立?事实上,因 $\boldsymbol{X}^{(0)}$ 的第 l 个分量对应于 $\boldsymbol{X}^{(1)}$ 的相应分量是零,即

$$x_l^{(0)}-\theta\beta_{l,m+t}=0$$

其中 $x_l^{(0)}$，θ 均不为零，根据 θ 规则（最小比值），$\beta_{l,m+t}\neq 0$。$\boldsymbol{X}^{(1)}$ 中的 m 个非零分量对应的 m 个列向量是 $\boldsymbol{P}_j(j=1,2,\cdots,m,j\neq l)$ 和 $\boldsymbol{P}_{m+t}$。若这组向量不是线性独立，则一定可以找到不全为零的数 α_j，使

$$\boldsymbol{P}_{m+t}=\sum_{\substack{j=1\\ j\neq l}}^{m}\alpha_j\boldsymbol{P}_j \tag{2-31}$$

成立。又因

$$\boldsymbol{P}_{m+t}=\sum_{j=1}^{m}\beta_{j,m+t}\boldsymbol{P}_j \tag{2-32}$$

将式(2-32)减式(2-31)得到

$$\sum_{\substack{j=1\\ j\neq l}}^{m}(\beta_{j,m+t}-\alpha_j)\boldsymbol{P}_j+\beta_{l,m+t}\boldsymbol{P}_l=0$$

由于上式中至少有 $\beta_{l,m+t}\neq 0$，所以上式表明 $\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_m$ 是线性相关，这与假设相矛盾。

由此可见，$\boldsymbol{X}^{(1)}$ 的 m 个非零分量对应的列向量 $\boldsymbol{P}_j(j=1,2,\cdots,m,j\neq l)$ 与 $\boldsymbol{P}_{m+t}$ 是线性独立的，即经过基变换得到的解是基可行解。实际上，从一个基可行解到另一个基可行解的变换，就是进行一次基变换。从几何意义上讲，就是从可行域的一个顶点转向另一个顶点（见图 2-2 解法）。

2.3.5　迭代（旋转运算）

上述讨论的基可行解的转换方法是用向量方程来描述，在实际计算时不太方便，因此采用系数矩阵法。现考虑以下形式的约束方程组

$$\begin{cases} x_1 \quad +a_{1,m+1}x_{m+1}+\cdots+a_{1k}x_k+\cdots+a_{1n}x_n=b_1\\ \quad x_2 \quad +a_{2,m+1}x_{m+1}+\cdots+a_{2k}x_k+\cdots+a_{2n}x_n=b_2\\ \qquad \ddots \\ \qquad x_l \quad +a_{l,m+1}x_{m+1}+\cdots+a_{lk}x_k+\cdots+a_{ln}x_n=b_l\\ \qquad\quad \ddots \\ \qquad\qquad x_m+a_{m,m+1}x_{m+1}+\cdots+a_{mk}x_k+\cdots+a_{mn}x_n=b_m \end{cases} \tag{2-33}$$

在一般线性规划问题的约束方程组中加入松弛变量或人工变量后，很容易得到上述形式。

设 $x_1,x_2,\cdots,x_m$ 为基变量，对应的系数矩阵是 $m\times m$ 单位阵 $\boldsymbol{I}$，它是可行基。令非基变量 $x_{m+1},x_{m+2},\cdots,x_n$ 为零，即可得到一个基可行解。若它不是最优解，则要另找一个使目标函数值增大的基可行解。这时从非基变量中确定 x_k 为换入变量。显然这时θ 为

$$\theta=\min_i\left(\frac{b_i}{a_{ik}}\middle| a_{ik}>0\right)=\frac{b_l}{a_{lk}}$$

在迭代过程中 θ 可表示为

$$\theta=\min_i\left(\frac{b_i'}{a_{ik}'}\middle| a_{ik}'>0\right)=\frac{b_l'}{a_{lk}'}$$

其中 b_i'，a_{ik}' 是经过迭代后对应于 b_i，a_{ik} 的元素值。

按 θ 规则确定 x_l 为换出变量，x_k，x_l 的系数列向量分别为

$$\boldsymbol{P}_k=\begin{bmatrix}a_{1k}\\a_{2k}\\\vdots\\a_{lk}\\\vdots\\a_{mk}\end{bmatrix};\quad \boldsymbol{P}_l=\begin{bmatrix}0\\\vdots\\1\\0\\\vdots\\0\end{bmatrix}\leftarrow \text{第 } l \text{ 个分量}$$

为了使 x_k 与 x_l 进行对换，须把 $\boldsymbol{P}_k$ 变为单位向量，这可以通过式(2-33)系数矩阵的增广矩阵进行初等变换来实现。

$$\begin{array}{c}
\begin{matrix} x_1 & \cdots & x_l & \cdots & x_m & x_{m+1} & \cdots & x_k & \cdots & x_n & b \end{matrix}\\
\left[\begin{array}{ccccccccc|c}
1 & & & & a_{1,m+1} & \cdots & a_{1k} & \cdots & a_{1n} & b_1\\
 & \ddots & & & & & & & & \\
 & & 1 & & a_{l,m+1} & \cdots & a_{lk} & \cdots & a_{ln} & b_l\\
 & & & \ddots & & & & & & \\
 & & & 1 & a_{m,m+1} & \cdots & a_{mk} & \cdots & a_{mn} & b_m
\end{array}\right]
\end{array} \tag{2-34}$$

变换的步骤是：

(1) 将增广矩阵式(2-34)中的第 l 行除以 a_{lk}，得到

$$\left(0,\cdots,0,\frac{1}{a_{lk}},0,\cdots,0,\frac{a_{l,m+1}}{a_{l,k}},\cdots,1,\cdots,\frac{a_{ln}}{a_{lk}}\middle|\frac{b_l}{a_{lk}}\right) \tag{2-35}$$

(2) 将式(2-34)中 x_k 列的各元素，除 a_{lk} 变换为 1 以外，其他都应变换为零。其他行的变换是将式(2-35)乘以 $a_{ik}(i\neq l)$后，从式(2-34)的第 i 行减去，得到新的第 i 行。

$$\left(0,\cdots,0,-\frac{a_{ik}}{a_{lk}},0,\cdots,0,a_{i,m+1}-\frac{a_{l,m+1}}{a_{lk}}a_{ik},\cdots,0,\cdots,a_{in}-\frac{a_{ln}}{a_{lk}}\cdot a_{ik}\middle|b_i-\frac{b_l}{a_{lk}}a_{ik}\right)$$

由此可得到变换后系数矩阵各元素的变换关系式：

$$a'_{ij}=\begin{cases}a_{ij}-\dfrac{a_{lj}}{a_{lk}}a_{ik} & (i\neq l)\\[2ex] \dfrac{a_{lj}}{a_{lk}} & (i=l)\end{cases};\quad b'_i=\begin{cases}b_i-\dfrac{a_{ik}}{a_{lk}}b_l & (i\neq l)\\[2ex] \dfrac{b_l}{a_{lk}} & (i=l)\end{cases}$$

a'_{ij}，b'_i 是变换后的新元素。

(3) 经过初等变换后的新增广矩阵是

$$\begin{array}{c}
\begin{matrix} x_1\cdots x_l \cdots x_m & x_{m+1}\cdots x_k\cdots x_n & b \end{matrix}\\
\left[\begin{array}{cccc|c}
1\cdots -\dfrac{a_{1k}}{a_{lk}}\cdots 0 & a'_{1,m+1}\cdots 0 & \cdots & a'_{1n} & b'_1\\
\vdots \qquad \vdots & \qquad\vdots & & & \vdots\\
0\cdots +\dfrac{1}{a_{lk}}\cdots 0 & a'_{l,m+1}\cdots 1 & \cdots & a'_{ln} & b'_l\\
\vdots \qquad \vdots & \qquad\vdots & & & \vdots\\
0\cdots -\dfrac{a_{mk}}{a_{lk}}\cdots 1 & a'_{m,m+1}\cdots 0 & \cdots & a'_{mn} & b'_m
\end{array}\right]
\end{array} \tag{2-36}$$

(4) 由式(2-36)中可以看到 $x_1, x_2, \cdots, x_k, \cdots, x_m$ 的系数列向量构成 $m \times m$ 单位矩阵，它是可行基，当非基变量 $x_{m+1}, \cdots, x_l, \cdots, x_n$ 为零时，就得到一个基可行解 $\boldsymbol{X}^{(1)}$。

$$\boldsymbol{X}^{(1)} = (b_1' \cdots b_{l-1}', 0, b_{l+1}' \cdots b_m', 0 \cdots b_k', 0 \cdots 0)^{\mathrm{T}}$$

在上述系数矩阵的变换中，元素 a_{lk} 称为主元素(也称为枢元素或旋转元(pivot element))，它所在列称为主元列，它所在行称为主元行。元素 a_{lk} 位置变换后为 1。

例 7 试用上述方法计算例 6 的一次基变换。

解 将例 6 的约束方程组的系数矩阵写成增广矩阵

$$\begin{array}{ccccc|c} x_1 & x_2 & x_3 & x_4 & x_5 & b \end{array}$$

$$\left(\begin{array}{ccccc|c} 1 & 2 & 1 & 0 & 0 & 8 \\ 4 & 0 & 0 & 1 & 0 & 16 \\ 0 & 4 & 0 & 0 & 1 & 12 \end{array}\right)$$

当以 x_3, x_4, x_5 为基变量，x_1, x_2 为非基变量，令 $x_1, x_2 = 0$，可得到一个基可行解

$$\boldsymbol{X}^{(0)} = (0, 0, 8, 16, 12)^{\mathrm{T}}$$

现用 x_2 去替换 x_5，于是将 x_3, x_4, x_2 的系数矩阵变换为单位矩阵，经变换后为

$$\begin{array}{ccccc|c} x_1 & x_2 & x_3 & x_4 & x_5 & b \end{array}$$

$$\left(\begin{array}{ccccc|c} 1 & 0 & 1 & 0 & -1/2 & 2 \\ 4 & 0 & 0 & 1 & 0 & 16 \\ 0 & 1 & 0 & 0 & 1/4 & 3 \end{array}\right)$$

令非基变量 $x_1, x_5 = 0$，得到新的基可行解

$$\boldsymbol{X}^{(1)} = (0, 3, 2, 16, 0)^{\mathrm{T}}$$

2.4 单纯形法的计算步骤

根据以上讨论的结果，将求解线性规划问题的单纯形法的计算步骤归纳如下。

2.4.1 单纯形表

为了便于理解计算关系，现设计一种计算表，称为单纯形表，其功能与增广矩阵相似，下面来建立这种计算表。

将式(2-22)与目标函数组成 $n+1$ 个变量，$m+1$ 个方程的方程组。

$$\begin{aligned} x_1 \qquad\quad & + a_{1m+1}x_{m+1} + \cdots + a_{1n}x_n = b_1 \\ x_2 \quad\; & + a_{2m+1}x_{m+1} + \cdots + a_{2n}x_n = b_2 \\ \ddots \quad & \\ x_m & + a_{mm+1}x_{m+1} + \cdots + a_{mn}x_n = b_m \\ -z + c_1x_1 + c_2x_2 + \cdots + c_mx_m & + c_{m+1}x_{m+1} + \cdots + c_nx_n = 0 \end{aligned}$$

为了便于迭代运算，可将上述方程组写成增广矩阵

$$
\begin{array}{c}
\begin{matrix} -z & x_1 & x_2 \cdots x_m & x_{m+1} \cdots x_n & b \end{matrix} \\
\left[\begin{array}{ccccc|c}
0 & 1 & 0 \cdots 0 & a_{1,m+1} \cdots a_{1n} & & b_1 \\
0 & 0 & 1 \cdots 0 & a_{2,m+1} \cdots a_{2n} & & b_2 \\
 & & \vdots & & & \\
0 & 0 & 0 \cdots 1 & a_{m,m+1} \cdots a_{mn} & & b_m \\
1 & c_1 & c_2 \cdots c_m & c_{m+1} \cdots c_n & & 0
\end{array}\right]
\end{array}
$$

若将 z 看做不参与基变换的基变量，它与 $x_1,x_2,\cdots,x_m$ 的系数构成一个基，这时可采用行初等变换将 $c_1,c_2,\cdots,c_m$ 变换为零，使其对应的系数矩阵为单位矩阵。得到

$$
\begin{array}{c}
\begin{matrix} -z & x_1 & x_2 \cdots x_m & x_{m+1} & \cdots & x_n & b \end{matrix} \\
\left[\begin{array}{cccccc|c}
0 & 1 & 0 \cdots 0 & a_{1,m+1} & \cdots & a_{1n} & b_1 \\
0 & 0 & 1 \cdots 0 & a_{2,m+1} & \cdots & a_{2n} & b_2 \\
\vdots & \vdots & \vdots \quad \vdots & \vdots & & \vdots & \vdots \\
0 & 0 & 0 \cdots 1 & a_{m,m+1} & \cdots & a_{mn} & b_m \\
1 & 0 & 0 \cdots 0 & c_{m+1} - \sum_{i=1}^{m} c_i a_{i,m+1} & \cdots & c_n - \sum_{i=1}^{m} c_i a_{in} & -\sum_{i=1}^{m} c_i b_i
\end{array}\right]
\end{array}
$$

可根据上述增广矩阵设计计算表，见表 2-2。

表 2-2

$c_j \to$			c_1	$\cdots$	c_m	c_{m+1}	$\cdots$	c_n	θ_i
C_B	X_B	b	x_1	$\cdots$	x_m	x_{m+1}	$\cdots$	x_n	
c_1	x_1	b_1	1	$\cdots$	0	$a_{1,m+1}$	$\cdots$	a_{1n}	θ_1
c_2	x_2	b_2	0		0	$a_{2,m+1}$		a_{2n}	θ_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
c_m	x_m	b_m	0	$\cdots$	1	$a_{m,m+1}$	$\cdots$	a_{mn}	θ_m
$c_j - z_j$			0	$\cdots$	0	$c_{m+1} - \sum_{i=1}^{m} c_i a_{i,m+1}$	$\cdots$	$c_n - \sum_{i=1}^{m} c_i a_{in}$	

X_B 列中填入基变量，这里是 $x_1,x_2,\cdots,x_m$；

C_B 列中填入基变量的价值系数，这里是 $c_1,c_2,\cdots,c_m$，它们是与基变量相对应的；

b 列中填入约束方程组右端的常数；

c_j 行中填入变量的价值系数 $c_1,c_2,\cdots,c_n$；

θ_i 列的数字是在确定换入变量后，按 θ 规则计算后填入；

最后一行称为检验数行，对应各非基变量 x_j 的检验数是

$$
c_j - \sum_{i=1}^{m} c_i a_{ij}, \quad j = 1,2,\cdots,n。
$$

表 2-2 称为初始单纯形表，每迭代一步构造一个新单纯形表。

2.4.2 计算步骤

(1) 根据数学模型确定初始可行基和初始基可行解，建立初始单纯形表。

(2) 计算各非基变量 x_j 的检验数是

$$\sigma_j = c_j - \sum_{i=1}^{m} c_i a_{ij}, \quad 若 \quad \sigma_j \leqslant 0, \quad j = m+1, \cdots, n$$

则已得到最优解，可终止计算。否则转入下一步。

(3) 在 $\sigma_j > 0, j = m+1, \cdots, n$ 中，若有某个 σ_k 对应 x_k 的系数列向量 $\boldsymbol{P}_k \leqslant 0$，则此问题属无界，终止计算。否则，转入下一步。

(4) 根据 $\max(\sigma_j > 0) = \sigma_k$，确定 x_k 为换入变量，按 θ 规则计算

$$\theta = \min\left(\frac{b_i}{a_{ik}} \middle| a_{ik} > 0\right) = \frac{b_l}{a_{lk}}$$

可确定 x_l 为换出变量，转入下一步。

(5) 以 a_{lk} 为主元素进行迭代（即用高斯消去法或称为旋转运算），把 x_k 所对应的列向量

$$\boldsymbol{P}_k = \begin{pmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{lk} \\ \vdots \\ a_{mk} \end{pmatrix} \xRightarrow{\text{变换为}} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \leftarrow 第\ l\ 行$$

将 $\boldsymbol{X}_B$ 列中的 x_l 换为 x_k，得到新的单纯形表。重复(2)～(5)，直到终止。

现用例 1 的标准型来说明上述计算步骤。

(1) 根据例 1 的标准型，取松弛变量 x_3, x_4, x_5 为基变量，它对应的单位矩阵为基。这就得到初始基可行解

$$\boldsymbol{X}^{(0)} = (0, 0, 8, 16, 12)^{\mathrm{T}}$$

将有关数字填入表中，得到初始单纯形表，见表 2-3。

表 2-3

$c_j \rightarrow$			2	3	0	0	0	θ_i
$\boldsymbol{C}_B$	$\boldsymbol{X}_B$	$\boldsymbol{b}$	x_1	x_2	x_3	x_4	x_5	
0	x_3	8	1	2	1	0	0	4
0	x_4	16	4	0	0	1	0	—
0	x_5	12	0	[4]	0	0	1	3
$c_j - z_j$			2	3	0	0	0	

表 2-3 中左上角的 c_j 是表示目标函数中各变量的价值系数。在 C_B 列填入初始基变量的价值系数,它们都为零,各非基变量的检验数为

$$\sigma_1=c_1-z_1=2-(0\times1+0\times4+0\times0)=2$$

$$\sigma_2=c_2-z_2=3-(0\times2+0\times0+0\times4)=3$$

(2) 因检验数都大于零,且 $\boldsymbol{P}_1,\boldsymbol{P}_2$ 有正分量存在,转入下一步;

(3) $\max(\sigma_1,\sigma_2)=\max(2,3)=3$,对应的变量 x_2 为换入变量,计算 θ

$$\theta=\min_i\left(\frac{b_i}{a_{i2}}\middle| a_{i2}>0\right)=\min\left(\frac{8}{2},-,\frac{12}{4}\right)=3$$

它所在行对应的 x_5 为换出变量。x_2 所在列和 x_5 所在行的交叉处[4]即为主元素。

(4) 以[4]为主元素进行旋转运算或迭代运算,即初等行变换,使 $\boldsymbol{P}_2$ 变换为 $(0,0,1)^{\mathrm{T}}$,在 $\boldsymbol{X}_B$ 列中用 x_2 替换 x_5,于是得到新表 2-4。

表 2-4

$c_j\rightarrow$			2	3	0	0	0	θ_i
C_B	X_B	b	x_1	x_2	x_3	x_4	x_5	
0	x_3	2	[1]	0	1	0	−1/2	2
0	x_4	16	4	0	0	1	0	4
3	x_2	3	0	1	0	0	1/4	—
c_j-z_j			2	0	0	0	−3/4	

$\boldsymbol{b}$ 列的数字是 $x_3=2,x_4=16,x_2=3$

于是得到新的基可行解 $\boldsymbol{X}^{(1)}=(0,3,2,16,0)^{\mathrm{T}}$

目标函数的取值 $z=9$

(5) 检查表 2-4 的所有 c_j-z_j,这时有 $c_1-z_1=2$;说明 x_1 应为换入变量。重复(2)~(4)的计算步骤,得表 2-5。

表 2-5

$c_j\rightarrow$			2	3	0	0	0	θ_i
C_B	X_B	b	x_1	x_2	x_3	x_4	x_5	
2	x_1	2	1	0	1	0	−1/2	—
0	x_4	8	0	0	−4	1	[2]	4
3	x_2	3	0	1	0	0	1/4	12
c_j-z_j			0	0	−2	0	1/4	

因表 2-5 中变量 x_5 列检验数为正,故继续迭代运算得表 2-6。

表 2-6

$c_j \to$			2	3	0	0	0	θ_i
C_B	X_B	b	x_1	x_2	x_3	x_4	x_5	
2	x_1	4	1	0	0	1/4	0	
0	x_5	4	0	0	−2	1/2	1	
3	x_2	2	0	1	1/2	−1/8	0	
$c_j - z_j$			0	0	−3/2	−1/8	0	

(6) 表 2-6 最后一行的所有检验数都已为负或零。这表示目标函数值已不可能再增大，于是得到最优解

$$X^* = X^{(3)} = (4,2,0,0,4)^T$$

目标函数值
$$z^* = 14$$

2.5 单纯形法的进一步讨论

2.5.1 人工变量法

在 2.3.2 节中提到，用添加人工变量法可以得到初始基可行解。这里加以讨论。

设线性规划问题的约束条件是

$$\sum_{j=1}^{n} P_j x_j = b$$

分别给每一个约束方程加入人工变量 $x_{n+1},\cdots,x_{n+m}$，得到

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n + x_{n+1} & = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \qquad + x_{n+2} & = b_2 \\ \qquad \cdots & \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \qquad\qquad + x_{n+m} & = b_m \\ x_1,\cdots,x_n \geqslant 0, x_{n+1},\cdots,x_{n+m} \geqslant 0 & \end{cases}$$

以 $x_{n+1},\cdots,x_{n+m}$ 为基变量，并可得到一个 $m \times m$ 单位矩阵。令非基变量 $x_1,\cdots,x_n$ 为零，便可得到一个初始基可行解

$$X^{(0)} = (0,0,\cdots,0,b_1,b_2,\cdots,b_m)^T$$

以下讨论如何解含有人工变量的线性规划问题。

1. 大 *M* 法

在一个线性规划问题的约束条件中加进人工变量后，要求人工变量对目标函数取值不受影响，为此假定人工变量在目标函数中的系数为($-M$)(M 为任意大的正数)，这样目标函数要实现最大化时，必须把人工变量从基变量换出。否则目标函数不可能实现最大化。

例 8 现有线性规划问题

$$\min z=-3x_1+x_2+x_3$$

$$\begin{cases} x_1-2x_2+x_3\leqslant 11 \\ -4x_1+x_2+2x_3\geqslant 3 \\ -2x_1+x_3=1 \\ x_1,x_2,x_3\geqslant 0 \end{cases}$$

试用大 M 法求解。

解 在上述问题的约束条件中加入松弛变量 x_4，剩余变量 x_5，人工变量 x_6,x_7，得到

$$\min z=-3x_1+x_2+x_3+0x_4+0x_5+Mx_6+Mx_7$$

$$\begin{cases} x_1-2x_2+x_3+x_4=11 \\ -4x_1+x_2+2x_3-x_5+x_6=3 \\ -2x_1+x_3+x_7=1 \\ x_1,x_2,x_3,x_4,x_5,x_6,x_7\geqslant 0 \end{cases}$$

这里 M 是一个任意大的正数。

用单纯形法进行计算时，见表 2-7。因本例是求 min，所以用所有 $c_j-z_j\geqslant 0$ 来判别目标函数是否实现了最小化。表 2-7 中的最终表表明已得到最优解是

$$x_1=4,x_2=1,x_3=9,x_4=x_5=x_6=x_7=0$$

目标函数 $z=-2$

表 2-7

c_j			-3	1	1	0	0	M	M	θ_i
C_B	X_B	b	x_1	x_2	x_3	x_4	x_5	x_6	x_7	
0	x_4	11	1	-2	1	1	0	0	0	11
M	x_6	3	-4	1	2	0	-1	1	0	3/2
M	x_7	1	-2	0	[1]	0	0	0	1	1
c_j-z_j			$-3+6M$	$1-M$	$1-3M$	0	M	0	0	
0	x_4	10	3	-2	0	1	0	0	-1	
M	x_6	1	0	[1]	0	0	-1	1	-2	1
1	x_3	1	-2	0	1	0	0	0	1	
c_j-z_j			-1	$1-M$	0	0	M	0	$3M-1$	
0	x_4	12	[3]	0	0	1	-2	2	-5	4
1	x_2	1	0	1	0	0	-1	1	-2	
1	x_3	1	-2	0	1	0	0	0	1	
c_j-z_j			-1	0	0	0	1	$M-1$	$M+1$	
-3	x_1	4	1	0	0	1/3	$-2/3$	2/3	$-5/3$	
1	x_2	1	0	1	0	0	-1	1	-2	
1	x_3	9	0	0	1	2/3	$-4/3$	4/3	$-7/3$	
c_j-z_j			0	0	0	1/3	1/3	$M-1/3$	$M-2/3$	

2. 两阶段法

下面介绍求解加入人工变量的线性规划问题的两阶段法。

第一阶段：不考虑原问题是否存在基可行解；给原线性规划问题加入人工变量，并构造仅含人工变量的目标函数和要求实现最小化。如

$$\min \omega = x_{n+1} + \cdots + x_{n+m} + 0x_1 + \cdots + 0x_n$$

$$\begin{cases} a_{11}x_1 + \cdots + a_{1n}x_n + x_{n+1} = b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n + x_{n+2} = b_2 \\ \qquad \cdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n + x_{n+m} = b_m \\ x_1, x_2, \cdots, x_{n+m} \geqslant 0 \end{cases}$$

然后用单纯形法求解上述模型，若得到 $\omega=0$，这说明原问题存在基可行解，可以进行第二段计算。否则原问题无可行解，应停止计算。

第二阶段：将第一阶段计算得到的最终表，除去人工变量。将目标函数行的系数，换原问题的目标函数系数，作为第二阶段计算的初始表。

各阶段的计算方法及步骤与 2.3 节单纯形法相同。下面举例说明。

例 9 线性规划问题

$$\min z = -3x_1 + x_2 + x_3$$

$$\begin{cases} x_1 - 2x_2 + x_3 \leqslant 11 \\ -4x_1 + x_2 + 2x_3 \geqslant 3 \\ -2x_1 + x_3 = 1 \\ x_1, \quad x_2, \quad x_3 \geqslant 0 \end{cases}$$

试用两阶段法求解。

解 先在上述线性规划问题的约束方程中加入人工变量，给出第一阶段的数学模型为

$$\min \omega = x_6 + x_7$$

$$\begin{cases} x_1 - 2x_2 + x_3 + x_4 = 11 \\ -4x_1 + x_2 + 2x_3 - x_5 + x_6 = 3 \\ -2x_1 + x_3 + x_7 = 1 \\ x_1, \ x_2, \ x_3, \ x_4, x_5, \ x_6, x_7 \geqslant 0 \end{cases}$$

这里 x_6, x_7 是人工变量。用单纯形法求解，见表 2-8。第一阶段求得的结果是 $\omega=0$，得到最优解是

$$x_1=0, x_2=1, x_3=1, x_4=12, x_5=x_6=x_7=0$$

因人工变量 $x_6=x_7=0$，所以 $(0,1,1,12,0)^{\mathrm{T}}$ 是这线性规划问题的基可行解。于是可以进行第二阶段运算。将第一阶段的最终表中的人工变量 x_6, x_7 的列删除，并在变量 x_1, x_2, x_3 的 c_j 处填入原问题的目标函数的系数。进行第二阶段计算，见表 2-9。

表 2-8

$c_j \to$			0	0	0	0	0	1	1	θ_i
C_B	X_B	b	x_1	x_2	x_3	x_4	x_5	x_6	x_7	
0	x_4	11	1	−2	1	1	0	0	0	11
1	x_6	3	−4	1	2	0	−1	1	0	3/2
1	x_7	1	−2	0	[1]	0	0	0	1	1
$c_j - z_j$			6	−1	−3	0	1	0	0	
0	x_4	10	3	−2	0	1	0	0	−1	—
1	x_6	1	0	[1]	0	0	−1	1	−2	1
0	x_3	1	−2	0	1	0	0	0	1	—
$c_j - z_j$			0	−1	0	0	1	0	3	
0	x_4	12	3	0	0	1	−2	2	−5	
0	x_2	1	0	1	0	0	−1	1	−2	
0	x_3	1	−2	0	1	0	0	0	1	
$c_j - z_j$			0	0	0	0	0	1	1	

表 2-9

c_j			−3	1	1	0	0	θ_i
C_B	X_B	b	x_1	x_2	x_3	x_4	x_5	
0	x_4	12	[3]	0	0	1	−2	4
1	x_2	1	0	1	0	0	−1	—
1	x_3	1	−2	0	1	0	0	—
$c_j - z_j$			−1	0	0	0	1	
−3	x_1	4	1	0	0	1/3	−2/3	
1	x_2	1	0	1	0	0	−1	
1	x_3	9	0	0	1	2/3	−4/3	
$c_j - z_j$			0	0	0	1/3	1/3	

从表 2-9 中得到最优解为 $x_1=4, x_2=1, x_3=9$，目标函数值 $z=-2$。

2.5.2 无可行解的判别规则

用单纯形法求解线性规划问题时，在添加人工变量后，才有可能出现无可行解。因为

人工变量是后加入到原约束条件中的虚拟变量，要求经过基的变换将它们从基变量中逐个替换出来。基变量中不再含有非零的人工变量，这表示原问题有解。若在最终表中当所有 $c_j-z_j\leqslant 0$，而在基变量中还有某个非零人工变量，这表示原问题无可行解。

2.5.3 退化

单纯形法计算中用 θ 规则确定换出变量时，有时存在两个以上相同的最小比值，这样在下一次迭代中就有一个或几个基变量等于零，这就出现退化解。这时换出变量 $x_l=0$，迭代后目标函数值不变，这时不同基表示为同一顶点。有人构造了一个特例，当出现退化时，进行多次迭代，而基从 $\boldsymbol{B}_1,\boldsymbol{B}_2,\cdots$，又返回到 $\boldsymbol{B}_1$，即出现计算过程的循环，便永远达不到最优解。

尽管计算过程的循环现象极少出现，但还是有可能的。如何解决这一问题？先后有人提出了"摄动法"，"字典序法"。1974 年由勃兰特(Bland)提出一种简便的规则，简称勃兰特规则：

(1) 选取 $c_j-z_j>0$ 中下标最小的非基变量 x_k 为换入变量，即

$$k=\min(j\mid c_j-z_j>0)$$

(2) 当按 θ 规则计算存在两个和两个以上最小比值时，选取下标最小的基变量为换出变量。

按勃兰特规则计算时，一定能避免出现循环。证明可见本篇参考资料[4]。

2.5.4 检验数的几种表示形式

本书以 $\max z=\boldsymbol{CX}$；$\boldsymbol{AX}=\boldsymbol{b},\boldsymbol{X}\geqslant 0$ 为标准型；以 $c_j-z_j\leqslant 0,(j=1,2,\cdots,n)$ 为最优解的判别准则，此外还有其他的形式。为了避免混淆，现将几种情况归纳如下。

设 $x_1,x_2,\cdots,x_m$ 为约束方程的基变量，于是可得

$$x_i=b_i-\sum_{j=m+1}^{n}a_{ij}x_j,\qquad i=1,2,\cdots,m$$

将它们代入目标函数后，可有两种表达形式

$$\begin{aligned} z&=\sum_{i=1}^{m}c_ib_i+\sum_{j=m+1}^{n}\left(c_j-\sum_{i=1}^{m}c_ia_{ij}\right)x_j \\ &=z_0+\sum_{j=m+1}^{n}(c_j-z_j)x_j \end{aligned}\tag{2-37}$$

或

$$z=\sum_{i=1}^{m}c_ib_i-\sum_{j=m+1}^{n}\left(\sum_{i=1}^{m}c_ia_{ij}-c_j\right)x_j$$

$$= z_0 - \sum_{j=m+1}^{n} (z_j - c_j) x_j \tag{2-38}$$

要求目标函数实现最大化时，若用式(2-37)来分析，就得到 $c_j - z_j \leqslant 0 (j=1,2,\cdots,n)$ 的判别准则。若用式(2-38)来分析，就得到 $z_j - c_j \geqslant 0 (j=1,2,\cdots,n)$ 的判别准则。

同样，在要求目标函数实现最小化时，可用式(2-37)或式(2-38)来分析，这时分别用 $c_j - z_j \geqslant 0$ 或 $z_j - c_j \leqslant 0 (j=1,2,\cdots,n)$ 来判别目标函数已达到最小。现将几种情况汇总于表 2-10。读者可按自己的习惯选用。

表 2-10

检验数	标准型	
	$\max z = CX$ $AX=b, X\geqslant 0$	$\min z = CX$ $AX=b, X\geqslant 0$
$c_j - z_j$	$\leqslant 0$	$\geqslant 0$
$z_j - c_j$	$\geqslant 0$	$\leqslant 0$

2.5.5 单纯形法小结

(1) 根据实际问题给出数学模型，列出初始单纯形表。进行标准化，见表 2-11。

表 2-11

变量	$x_j \geqslant 0$		不需要处理
	$x_j \leqslant 0$		令 $x_j' = -x_j$；$x_j' \geqslant 0$
	x_j 无约束		令 $x_j = x_j' - x_j''$；$x_j', x_j'' \geqslant 0$
约束条件	$b \geqslant 0$		不需要处理
	$b < 0$		约束条件两端同乘 -1
	$\leqslant$		加松弛变量 x_{si}
	$=$		加人工变量 x_{ai}
	$\geqslant$		减去剩余(松弛)变量 x_{si}，加人工变量 x_{ai}
目标函数	$\max z$		不需要处理
	$\min z$		令 $z' = -z$，求 $\max z'$
	加入变量的系数	松弛变量 x_{si}	0
		人工变量 x_{ai}	$-M$

分别以每个约束条件中的松弛变量或人工变量为基变量，列出初始单纯形表。

(2) 对目标函数求 max 的线性规划问题,用单纯形法计算步骤的框图见图 2-9。

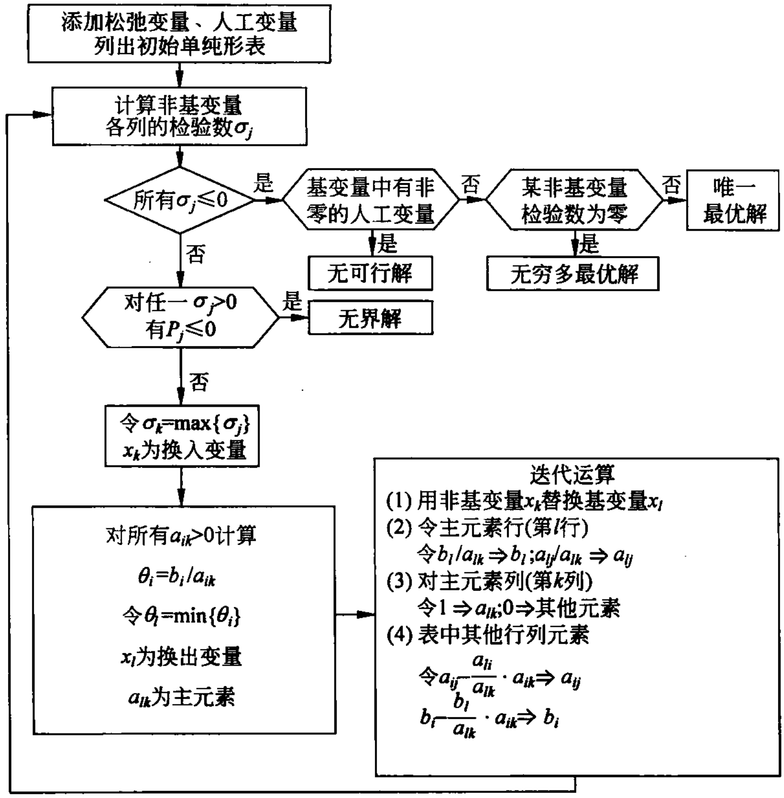


图 2-9

2.6 应用举例

一般来讲,一个经济、管理问题凡满足以下条件时,才能建立线性规划的模型。

(1) 要求解问题的目标函数能用数值指标来反映,且为线性函数;

(2) 存在着多种方案及有关数据;

(3) 要求达到的目标是在可以量化的,并且有足够数据的一定约束条件下实现的,这些约束条件可用线性等式或不等式来描述。

下面举例说明线性规划在经济管理等方面的应用。

例 10 合理利用线材问题。现要做 100 套钢架,每套用长为 2.9m,2.1m 和 1.5m 的元钢各一根。已知原料长 7.4m,问应如何下料,使用的原材料最省。

解 最简单做法是,在每一根原材料上截取 2.9m,2.1m 和 1.5m 的元钢各一根组成一套,每根原材料剩下料头 0.9m。为了做 100 套钢架,需用原材料 100 根,有 90m 料头。

若改为用套裁,这可以节约原材料。下面有几种套裁方案,都可以考虑采用。见表2-12。

表 2-12

长度/m	下料根数				
	方案				
	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ
2.9	1	2		1	
2.1			2	2	1
1.5	3	1	2		3
合计	7.4	7.3	7.2	7.1	6.6
料头	0	0.1	0.2	0.3	0.8

为了得到100套钢架,需要混合使用各种下料方案。设按Ⅰ方案下料的原材料根数为 x_1,Ⅱ方案为 x_2,Ⅲ方案为 x_3,Ⅳ方案为 x_4,Ⅴ方案为 x_5。根据表2-12的方案,可列出以下数学模型

$$\min z = 0x_1 + 0.1x_2 + 0.2x_3 + 0.3x_4 + 0.8x_5$$

$$\begin{cases} x_1 + 2x_2 + x_4 = 100 \\ 2x_3 + 2x_4 + x_5 = 100 \\ 3x_1 + x_2 + 2x_3 + 3x_5 = 100 \\ x_1, x_2, x_3, x_4, x_5 \geqslant 0 \end{cases}$$

在以上约束条件中加入人工变量 x_6, x_7, x_8;然后用表2-13进行计算。

由计算得到最优下料方案是:按Ⅰ方案下料30根;Ⅱ方案下料10根;Ⅳ方案下料50根。即需90根原材料可以制造100套钢架。

例11 配料问题

某工厂要用三种原材料 C、P、H 混合调配出三种不同规格的产品 A、B、D。已知产品的规格要求,产品单价,每天能供应的原材料数量及原材料单价,分别见表2-14和表2-15。该厂应如何安排生产,使利润收入为最大?

解 如以 A_C 表示产品 A 中 C 的成分,A_P 表示产品 A 中 P 的成分,依此类推。

根据表2-14有:

$$A_C \geqslant \frac{1}{2}A, A_P \leqslant \frac{1}{4}A, B_C \geqslant \frac{1}{4}B, B_P \leqslant \frac{1}{2}B \tag{2-39}$$

这里

$$\begin{aligned} A_C + A_P + A_H &= A \\ B_C + B_P + B_H &= B \end{aligned} \tag{2-40}$$

将式(2-40)逐个代入式(2-39)并整理得到

$$-\frac{1}{2}A_C + \frac{1}{2}A_P + \frac{1}{2}A_H \leqslant 0$$

表　2-13

$c_j \to$			0	−0.1	−0.2	−0.3	−0.8	−M	−M	−M	θ_i
C_B	X_B	b	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	
−M	x_6	100	1	2	0	1	0	1	0	0	$\frac{100}{1}$
−M	x_7	100	0	0	2	2	1	0	1	0	—
−M	x_8	100	[3]	1	2	0	3	0	0	1	$\frac{100}{3}$
$c_j - z_j$			4M	−0.1+3M	−0.2+4M	−0.3+3M	−0.8+4M	0	0	0	
−M	x_6	200/3	0	5/3	−2/3	1	−1	1	0	−1/3	$\frac{200}{3}$
−M	x_7	100	0	0	2	[2]	1	0	1	0	$\frac{100}{2}$
0	x_1	100/3	1	1/3	2/3	0	1	0	0	1/3	—
$c_j - z_j$			0	−0.1+5/3M	−0.2+4/3M	−0.3+3M	−0.8	0	0	−4/3M	
−M	x_6	50/3	0	[5/3]	−5/3	0	−3/2	1	−1/2	−1/3	$\frac{150}{15}$
−0.3	x_4	50	0	0	1	1	1/2	0	1/2	0	—
0	x_1	100/3	1	1/3	2/3	0	1	0	0	1/3	$\frac{100}{1}$
$c_j - z_j$			0	−0.1+5/3M	0.1−5/3M	0	−0.65−3/2M	0	0.15−3/2M	−4/3M	
0.1	x_2	10	0	1	−1	0	−9/10	3/5	−3/10	−1/5	
−0.3	x_4	50	0	0	1	1	1/2	0	1/2	0	
0	x_1	30	1	0	1	0	13/10	−1/5	1/10	2/5	
$c_j - z_j$			0	0	0	0	−0.74	−M+0.06	−M+0.12	−M−0.02	

本例中存在多重最优解，请读者自检。

表　2-14

产品名称	规格要求	单价/(元/kg)
A	原材料 *C* 不少于 50% 原材料 *P* 不超过 25%	50
B	原材料 *C* 不少于 25% 原材料 *P* 不超过 50%	35
D	不限	25

$$-\frac{1}{4}A_C+\frac{3}{4}A_P-\frac{1}{4}A_H\leqslant 0$$

$$-\frac{3}{4}B_C+\frac{1}{4}B_P+\frac{1}{4}B_H\leqslant 0$$

$$-\frac{1}{2}B_C+\frac{1}{2}B_P-\frac{1}{2}B_H\leqslant 0$$

表 2-15

原材料名称	每天最多供应量/kg	单价/(元/kg)
C	100	65
P	100	25
H	60	35

表 2-15 表明这些原材料供应数量的限额。加入到产品 A、B、D 的原材料 C 总量每天不超过 100kg，P 的总量不超过 100kg，H 总量不超过 60kg。由此

$$A_C+B_C+D_C\leqslant 100$$

$$A_P+B_P+D_P\leqslant 100$$

$$A_H+B_H+D_H\leqslant 60$$

在约束条件中共有 9 个变量，为计算和叙述方便，分别用 $x_1,\cdots,x_9$ 表示。令

$$x_1=A_C\qquad x_2=A_P\qquad x_3=A_H$$

$$x_4=B_C\qquad x_5=B_P\qquad x_6=B_H$$

$$x_7=D_C\qquad x_8=D_P\qquad x_9=D_H$$

由此约束条件可表示为

$$\begin{cases}-\frac{1}{2}x_1+\frac{1}{2}x_2+\frac{1}{2}x_3\leqslant 0\\ -\frac{1}{4}x_1+\frac{3}{4}x_2-\frac{1}{4}x_3\leqslant 0\\ -\frac{3}{4}x_4+\frac{1}{4}x_5+\frac{1}{4}x_6\leqslant 0\\ -\frac{1}{2}x_4+\frac{1}{2}x_5-\frac{1}{2}x_6\leqslant 0\\ x_1+x_4+x_7\leqslant 100\\ x_2+x_5+x_8\leqslant 100\\ x_3+x_6+x_9\leqslant 60\\ x_1,x_2,\cdots,x_9\geqslant 0\end{cases}$$

我们的目的是使利润最大，即产品价格减去原材料的价格为最大。

产品价格为　　$50(x_1+x_2+x_3)$——产品 A

$35(x_4+x_5+x_6)$——产品 B

$$25(x_7+x_8+x_9)——产品 D$$

原材料价格为

$$65(x_1+x_4+x_7)——原材料 C$$
$$25(x_2+x_5+x_8)——原材料 P$$
$$35(x_3+x_6+x_9)——原材料 H$$

目标函数

$$\begin{aligned}\max z &=50(x_1+x_2+x_3)+35(x_4+x_5+x_6)+25(x_7+x_8+x_9)\\&\quad -65(x_1+x_4+x_7)-25(x_2+x_5+x_8)-35(x_3+x_6+x_9)\\&=-15x_1+25x_2+15x_3-30x_4+10x_5-40x_7-10x_9\end{aligned}$$

为了得到初始解，在约束条件中加入松弛变量 $x_{10}\sim x_{16}$，得到数学模型

$$\begin{aligned}\max z=&-15x_1+25x_2+15x_3-30x_4+10x_5-40x_7-10x_9\\&+0\cdot(x_{10}+x_{11}+x_{12}+x_{13}+x_{14}+x_{15}+x_{16})\end{aligned}$$

$$\begin{cases}-\frac{1}{2}x_1+\frac{1}{2}x_2+\frac{1}{2}x_3+x_{10}=0\\ -\frac{1}{4}x_1+\frac{3}{4}x_2-\frac{1}{4}x_3+x_{11}=0\\ -\frac{3}{4}x_4+\frac{1}{4}x_5+\frac{1}{4}x_6+x_{12}=0\\ -\frac{1}{2}x_4+\frac{1}{2}x_5-\frac{1}{2}x_6+x_{13}=0\\ x_1+x_4+x_7+x_{14}=100\\ x_2+x_5+x_8+x_{15}=100\\ x_3+x_6+x_9+x_{16}=60\\ x_i\geqslant 0, i=1,2,\cdots,16\end{cases}$$

上述数学模型，可用单纯形法计算，经过 4 次迭代，获得最优解为：$x_1=100$，$x_2=50$，$x_3=50$；这表示需要用原料 C 为 100kg，P 为 50kg，H 为 50kg 构成产品 A，即每天只生产产品 A 为 200kg。从最终计算表中得到，总利润是 $z=500$ 元/天。

例 12 某快递公司下设一个快件分拣部，处理每天到达和外寄的快件。根据统计资料及经验预测，每天各时段快件数量如表 2-16 所示。

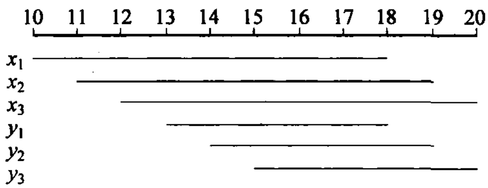
表 2-16

时 段	到达快件数	时 段	到达快件数
10:00 前	5 000	14:00～15:00	3 000
10:00～11:00	4 000	15:00～16:00	4 000
11:00～12:00	3 000	16:00～17:00	4 500
12:00～13:00	4 000	17:00～18:00	3 500
13:00～14:00	2 500	18:00～19:00	2 500

快件分拣由机器操作,分拣效率为每台 500 体/h,每台机器操作时需要配一名职工,共有 11 台机器。分拣部一部分是全日制职工,上班时间分别为 10:00～18:00,11:00～19:00 和 12:00～20:00,每人每天工资 150 元;另一部分是非全日制职工,每天上班 5 小时,分别为 13:00～18:00,14:00～19:00,15:00～20:00,每人每天工资 80 元。快件处理的规则是从每个整点起可处理该整点前到达的快件,例如从 11:00 起可处理 10:00 前和 10:00～11:00 之间到达的,13:00 起可处理所有这之前到达的等。因快件有时间性要求,凡是 12:00 前到达的快件必须在 14:00 以前处理完;15:00 以前到达的,必须在 17:00 以前处理完;全部快件在当天 20:00 以前处理完。

问该分拣部要完成快件处理任务,应设多少名全日制及非全日制职工,并使总的工资支出为最少。

解 设 x_1、x_2、x_3 分别为从 10:00～18:00、11:00～19:00、12:00～20:00 上班的全日制职工,y_1、y_2、y_3 分别为从 13:00～18:00、14:00～19:00、15:00～20:00 上班的非全日制职工。可用下图表示职工上班的时段。用下图表示职工上班时间。



根据题意,可列出目标函数和约束条件如下

$$\min z = 150(x_1 + x_2 + x_3) + 80(y_1 + y_2 + y_3)$$

$$\begin{cases}
500x_1 \leqslant 5\,000 & ①\\
1\,000x_1 + 500x_2 \leqslant 9\,000 & ②\\
1\,500x_1 + 1\,000x_2 + 500x_3 \leqslant 12\,000 & ③\\
2\,000x_1 + 1\,500x_2 + 1\,000x_3 + 500y_1 \leqslant 16\,000 & ④\\
2\,500x_1 + 2\,000x_2 + 1\,500x_3 + 1\,000y_1 + 500y_2 \leqslant 18\,500 & ⑤\\
3\,000x_1 + 2\,500x_2 + 2\,000x_3 + 1\,500y_1 + 1\,000y_2 + 500y_3 \leqslant 21\,500 & ⑥\\
3\,500x_1 + 3\,000x_2 + 2\,500x_3 + 2\,000y_1 + 1\,500y_2 + 1\,000y_3 \leqslant 25\,500 & ⑦\\
4\,000x_1 + 3\,500x_2 + 3\,000x_3 + 2\,500y_1 + 2\,000y_2 + 1\,500y_3 \leqslant 30\,000 & ⑧\\
4\,000x_1 + 4\,000x_2 + 3\,500x_3 + 2\,500y_1 + 2\,500y_2 + 2\,000y_3 \leqslant 33\,500 & ⑨\\
4\,000x_1 + 4\,000x_2 + 4\,000x_3 + 2\,500y_1 + 2\,500y_2 + 2\,500y_3 \geqslant 36\,000 & ⑩\\
2\,000x_1 + 1\,500x_2 + 1\,000x_3 + 500y_1 \geqslant 12\,000 & ⑪\\
3\,500x_1 + 3\,000x_2 + 2\,500x_3 + 2\,000y_1 + 1\,500y_2 + 1\,000y_3 \geqslant 21\,500 & ⑫\\
x_1 + x_2 + x_3 + y_1 + y_2 + y_3 \leqslant 11 & ⑬\\
x_j \geqslant 0(j = 1,\cdots 3),\ y_j \geqslant 0(j = 1,\cdots 3)
\end{cases}$$

约束条件①～⑨为各时段内投入两类职工数和可处理该时段前到达的快件数,⑩～

⑫为快件处理时限的要求，⑬为分拣机器的限制。

本题求解的结果有两个：

(1) $x_1=2,x_2=2,x_3=5,y_1=0,y_2=0,y_3=0$；

(2) $x_1=3,x_2=0,x_3=6,y_1=0,y_2=0,y_3=0$.

总工资支出均为 1 350 元/天。

例 13　连续投资问题

某部门在今后五年内考虑给下列项目投资，已知：

项目 A，从第 1—第 4 年每年年初需要投资，并于次年年末回收本利 115%；

项目 B，第 3 年年初需要投资，到第 5 年年末能回收本利 125%，但规定最大投资额不超过 4 万元；

项目 C，第 2 年年初需要投资，到第 5 年年末能回收本利 140%，但规定最大投资额不超过 3 万元；

项目 D，5 年内每年年初可购买公债，于当年年末归还，并加利息 6%。

该部门现有资金 10 万元，问它应如何确定给这些项目每年的投资额，使到第 5 年年末拥有的资金的本利总额为最大？

解

(1) 确定决策变量

这是一个连续投资问题，与时间有关。但这里设法用线性规划方法，静态地处理，以 $x_{iA},x_{iB},x_{iC},x_{iD}(i=1,2,\cdots,5)$ 分别表示第 i 年年初给项目 A,B,C,D 的投资额。它们都是待定的未知变量。根据给定的条件，将变量列于表 2-17 中。

表　2-17

项　目	第 1 年	第 2 年	第 3 年	第 4 年	第 5 年
A	x_{1A}	x_{2A}	x_{3A}	x_{4A}	
B			x_{3B}		
C		x_{2C}			
D	x_{1D}	x_{2D}	x_{3D}	x_{4D}	x_{5D}

(2) 投资额应等于手中拥有的资金额

由于项目 D 每年都可以投资，并且当年年末即能回收本息。所以该部门每年应把资金全部投出去，手中不应当有剩余的呆滞资金。因此

第 1 年：该部门年初拥有 100 000 元，所以有

$$x_{1A}+x_{1D}=100\ 000$$

第 2 年：因第 1 年给项目 A 的投资要到第 2 年年末才能回收。所以该部门在第 2 年年初拥有资金额仅为项目 D 在第一年回收的本息 $x_{1D}(1+6\%)$。于是第二年的投资分配是

$$x_{2A}+x_{2C}+x_{2D}=1.06x_{1D}$$

第 3 年：第 3 年年初的资金额是从项目 A 第 1 年投资及项目 D 第 2 年投资中回收的本利总和：$x_{1A}(1+15\%)$ 及 $x_{2D}(1+6\%)$。于是第 3 年的资金分配为

$$x_{3A}+x_{3B}+x_{3D}=1.15x_{1A}+1.06x_{2D}$$

第 4 年：同以上分析，可得

$$x_{4A}+x_{4D}=1.15x_{2A}+1.06x_{3D}$$

第 5 年：

$$x_{5D}=1.15x_{3A}+1.06x_{4D}$$

此外，由于对项目 B、C 的投资有限额的规定，即

$$x_{3B}\leqslant 40\,000$$
$$x_{2C}\leqslant 30\,000$$

(3) 目标函数

问题是要求在第五年年末该部门手中拥有的资金额达到最大，与 5 年年末资金有关的变量是：$x_{4A}, x_{3B}, x_{2C}, x_{5D}$，因此这个目标函数可表示为

$$\max\ z=1.15x_{4A}+1.40x_{2C}+1.25x_{3B}+1.06x_{5D}$$

(4) 数学模型

经过以上分析，这个与时间有关的投资问题可以用以下线性规划模型来描述

$$\max\ z=1.15x_{4A}+1.40x_{2C}+1.25x_{3B}+1.06x_{5D}$$

满足

$$\begin{cases} x_{1A}+x_{1D} & =100\,000 \\ \quad -1.06x_{1D}+x_{2A}+x_{2C}+x_{2D} & =0 \\ -1.15x_{1A}-1.06x_{2D}+x_{3A}+x_{3B}+x_{3D} & =0 \\ -1.15x_{2A}-1.06x_{3D}+x_{4A}+x_{4D} & =0 \\ -1.15x_{3A}-1.06x_{4D}+x_{5D} & =0 \\ x_{2C} & \leqslant 30\,000 \\ x_{3B} & \leqslant 40\,000 \\ x_{iA}, x_{iB}, x_{iC}, x_{iD}\geqslant 0 \quad i=1,2,\cdots,5 & \end{cases}$$

(5) 用两阶段单纯形法计算结果得到

第 1 年：$x_{1A}=34\,783$ 元，$x_{1D}=65\,217$ 元

第 2 年：$x_{2A}=39\,130$ 元，$x_{2C}=30\,000$ 元，$x_{2D}=0$

第 3 年：$x_{3A}=0$，$x_{3B}=40\,000$ 元，$x_{3D}=0$

第 4 年：$x_{4A}=45\,000$ 元，$x_{4D}=0$

第 5 年：$x_{5D}=0$

另一个投资方案：

第 1 年：$x_{1A}=71\,698$ 元，$x_{1D}=28\,300$ 元

第 2 年：$x_{2A}=0$ 元，$x_{2C}=30\,000$ 元，$x_{2D}=0$ 元

第 3 年：$x_{3A}=42\,453$ 元，$x_{3B}=40\,000$ 元，$x_{3D}=0$ 元

第 4 年：$x_{4A}=0$ 元，$x_{4D}=0$ 元

第 5 年：$x_{5D}=48\ 820$ 元。

还可以有其他的方案。

到第五年年末该部门拥有资金总额为 143 750 元，即赢利 43.75%。

习题

2.1 用图解法求解下列线性规划问题，并指出问题是具有唯一最优解、无穷多最优解、无界解还是无可行解？

(1) $\max z=x_1+3x_2$

$$\begin{cases}5x_1+10x_2\leqslant 50\\ x_1+x_2\geqslant 1\\ x_2\leqslant 4\\ x_1,x_2\geqslant 0\end{cases}$$

(2) $\min z=x_1+1.5x_2$

$$\begin{cases}x_1+3x_2\geqslant 3\\ x_1+x_2\geqslant 2\\ x_1,x_2\geqslant 0\end{cases}$$

(3) $\max z=2x_1+2x_2$

$$\begin{cases}x_1-x_2\geqslant -1\\ -0.5x_1+x_2\leqslant 2\\ x_1,x_2\geqslant 0\end{cases}$$

(4) $\max z=x_1+x_2$

$$\begin{cases}x_1-x_2\geqslant 0\\ 3x_1-x_2\leqslant -3\\ x_1,x_2\geqslant 0\end{cases}$$

2.2 将下列线性规划问题变换成标准型，并列出初始单纯形表。

(1) $\min z=-3x_1+4x_2-2x_3+5x_4$

$$\begin{cases}4x_1-x_2+2x_3-x_4=-2\\ x_1+x_2+3x_3-x_4\leqslant 14\\ -2x_1+3x_2-x_3+2x_4\geqslant 2\\ x_1,x_2,x_3\geqslant 0,x_4\text{ 无约束}\end{cases}$$

(2) $\max s=z_k/p_k$

$$\begin{cases}z_k=\sum_{i=1}^{n}\sum_{k=1}^{m}\alpha_{ik}x_{ik}\\ \sum_{k=1}^{m}-x_{ik}=-1(i=1,\cdots,n)\\ x_{ik}\geqslant 0(i=1,\cdots,n;\ k=1,\cdots,m)\end{cases}$$

2.3 在下面的线性规划问题中找出满足约束条件的所有基解。指出哪些是基可行解，并代入目标函数，确定哪一个是最优解。

(1) $\max z=2x_1+3x_2+4x_3+7x_4$

$$\begin{cases}2x_1+3x_2-x_3-4x_4=8\\ x_1-2x_2+6x_3-7x_4=-3\\ x_1,x_2,x_3,x_4\geqslant 0\end{cases}$$

(2) $\max z=5x_1-2x_2+3x_3-6x_4$

$$\begin{cases}x_1+2x_2+3x_3+4x_4=7\\ 2x_1+x_2+x_3+2x_4=3\\ x_1,x_2,x_3,x_4\geqslant 0\end{cases}$$

2.4 用单纯形法求解习题 2.1 中线性规划问题(1)～(4)，要求：

(1) 分别验证在 2.3.3 节和 2.5.2 节给出的问题具有唯一最优解、无穷多最优解、无界解、及无解时的判别标准。

(2) 当问题有可行解时，指出单纯形法迭代的每一步相当于图形上哪一个顶点。

2.5 以

$$\max z = 2x_1 + x_2$$
$$\begin{cases} 3x_1 + 5x_2 \leqslant 15 \\ 6x_1 + 2x_2 \leqslant 24 \\ x_1, x_2 \geqslant 0 \end{cases}$$

为例用图解法，具体说明当目标函数中变量的系数怎样改变时，使满足约束条件的可行域的每一个顶点，都有可能使目标函数值达到最优。

2.6 分别用单纯形法中的大 M 法和两阶段法求解下述线性规划问题，并指出属哪一类解。

(1) $\max z = 2x_1 + 3x_2 - 5x_3$

$$\begin{cases} x_1 + x_2 + x_3 = 7 \\ 2x_1 - 5x_2 + x_3 \geqslant 10 \\ x_1, \quad x_2, \quad x_3 \geqslant 0 \end{cases}$$

(2) $\min z = 2x_1 + 3x_2 + x_3$

$$\begin{cases} x_1 + 4x_2 + 2x_3 \geqslant 8 \\ 3x_1 + 2x_2 \geqslant 6 \\ x_1, \quad x_2, \quad x_3 \geqslant 0 \end{cases}$$

2.7 求下述线性规划问题目标函数 z 的上界 $\bar{z}^*$ 和下界 $\underline{z}^*$

$$\max z = c_1 x_1 + c_2 x_2$$
$$\begin{cases} a_{11} x_1 + a_{12} x_2 \leqslant b_1 \\ a_{21} x_1 + a_{22} x_2 \leqslant b_2 \\ x_1, x_2 \geqslant 0 \end{cases}$$

其中：$1 \leqslant c_1 \leqslant 3$，$4 \leqslant c_2 \leqslant 6$，$8 \leqslant b_1 \leqslant 12$，$10 \leqslant b_2 \leqslant 14$，$-1 \leqslant a_{11} \leqslant 3$，$2 \leqslant a_{12} \leqslant 5$，$2 \leqslant a_{21} \leqslant 4$，$4 \leqslant a_{22} \leqslant 6$

2.8 表 2-18 是某求极大化线性规划问题计算得到的单纯形表。表中无人工变量，a_1、a_2、a_3、d、c_1、c_2 为待定常数。试说明这些常数分别取何值时，以下结论成立。

(1) 表中解为唯一最优解；

(2) 表中解为最优解，但存在无穷多最优解；

(3) 该线性规划问题具有无界解；

(4) 表中解非最优，为对解改进，换入变量为 x_1，换出变量为 x_6。

表 2-18

基 b	x_1	x_2	x_3	x_4	x_5	x_6
x_3 d	4	a_1	1	0	a_2	0
x_4 2	−1	−3	0	1	−1	0
x_6 3	a_3	−5	0	0	−4	1
$c_j - z_j$	c_1	c_2	0	0	−3	0

2.9 某昼夜服务的公交线路每天各时间区段内所需司机和乘务人员数如表 2-19：

表 2-19

班次	时 间	所需人数
1	6:00～10:00	60
2	10:00～14:00	70
3	14:00～18:00	60
4	18:00～22:00	50
5	22:00～2:00	20
6	2:00～6:00	30

设司机和乘务人员分别在各时间区段一开始时上班，并连续工作八小时，问该公交线路至少配备多少名司机和乘务人员。列出这个问题的线性规划模型。

2.10 某糖果厂用原料 A、B、C 加工成三种不同牌号的糖果甲、乙、丙。已知各种牌号糖果中 A、B、C 含量，原料成本，各种原料的每月限制用量，三种牌号糖果的单位加工费及售价如表 2-20 表示。

表 2-20

原 料	甲	乙	丙	原料成本/(元/kg)	每月限制用量/kg
A	≥60%	≥15%		2.00	2 000
B				1.50	2 500
C	≤20%	≤60%	≤50%	1.00	1 200
加工费/(元/kg)	0.50	0.40	0.30		
售价/(元/kg)	3.40	2.85	2.25		

问该厂每月应生产这三种牌号糖果各多少千克，使该厂获利最大？试建立这个问题的线性规划的数学模型。

2.11 某厂生产三种产品Ⅰ，Ⅱ，Ⅲ。每种产品要经过 A，B 两道工序加工。设该厂有两种规格的设备能完成 A 工序，它们以 A_1，A_2 表示；有三种规格的设备能完成 B 工序，它们以 B_1，B_2，B_3 表示。产品Ⅰ可在 A，B 任何一种规格设备上加工。产品Ⅱ可在任何规格的 A 设备上加工，但完成 B 工序时，只能在 B_1 设备上加工；产品Ⅲ只能在 A_2 与 B_2 设备上加工。已知在各种机床设备的单件工时，原材料费，产品销售价格，各种设备有效台时以及满负荷操作时机床设备的费用如表 2-21，要求安排最优的生产计划，使该厂利润最大。

表 2-21

设 备	产 品			设备有效台时/台时	满负荷时的设备费用/元
	Ⅰ	Ⅱ	Ⅲ		
A_1	5	10		6 000	300
A_2	7	9	12	10 000	321
B_1	6	8		4 000	250
B_2	4		11	7 000	783
B_3	7			4 000	200
原料费/(元/件)	0.25	0.35	0.50		
单价/(元/件)	1.25	2.00	2.80		

CHAPTER 3
第3章

对偶理论和灵敏度分析

3.1 单纯形法的矩阵描述

现在用矩阵描述单纯形法的计算过程。它将有助于对单纯形法的理解，以及学习对偶理论和灵敏度分析。

设线性规划问题：$\max z=\boldsymbol{CX}$；$\boldsymbol{AX}\leqslant \boldsymbol{b}$；$\boldsymbol{X}\geqslant 0$。在加入松弛变量 $\boldsymbol{X}_S=(x_{s_1},x_{s_2},\cdots,x_{s_n})^{\mathrm{T}}$ 以后，它对应的目标函数中系数 $\boldsymbol{C}_S=0$，得到标准型

$$\max z=\boldsymbol{CX}+\boldsymbol{C}_S\boldsymbol{X}_S;\quad \boldsymbol{AX}+\boldsymbol{IX}_S=\boldsymbol{b};\quad \boldsymbol{X},\boldsymbol{X}_S\geqslant 0 \tag{3-1}$$

这里的 $\boldsymbol{I}$ 是 $m\times m$ 单位矩阵。若以 $\boldsymbol{X}_S$ 为基变量，这时可标记成 $\boldsymbol{X}_B$。其对应的单位矩阵便是基矩阵 $\boldsymbol{B}$，这时将系数矩阵 $(\boldsymbol{A},\boldsymbol{I})$ 分为 $(\boldsymbol{B},\boldsymbol{N})$ 两块。$\boldsymbol{N}$ 是非基变量的系数矩阵。相应的决策变量被分为 $\boldsymbol{X}=\begin{pmatrix}\boldsymbol{X}_B\\ \boldsymbol{X}_N\end{pmatrix}$；同时将目标函数的系数 $(\boldsymbol{C},\boldsymbol{C}_S)$ 分别对应于基变量和非基变量，并记作 $(\boldsymbol{C}_B,\boldsymbol{C}_N)$。这时不论是初始状态，还是迭代后的关系式都可以表述为

目标函数：
$$\max z=(\boldsymbol{C}_B,\boldsymbol{C}_N)\begin{pmatrix}\boldsymbol{X}_B\\ \boldsymbol{X}_N\end{pmatrix}=\boldsymbol{C}_B\boldsymbol{X}_B+\boldsymbol{C}_N\boldsymbol{X}_N \tag{3-2}$$

约束条件：
$$(\boldsymbol{B},\boldsymbol{N})\begin{pmatrix}\boldsymbol{X}_B\\ \boldsymbol{X}_N\end{pmatrix}=\boldsymbol{BX}_B+\boldsymbol{NX}_N=\boldsymbol{b} \tag{3-3}$$

非负条件：
$$\boldsymbol{X}_B,\boldsymbol{X}_N\geqslant 0 \tag{3-4}$$

将式(3-3)移项后，得 $\boldsymbol{BX}_B=\boldsymbol{b}-\boldsymbol{NX}_N$，然后给等式两边左乘 $\boldsymbol{B}^{-1}$ 后，可解得

$$\boldsymbol{X}_B=\boldsymbol{B}^{-1}\boldsymbol{b}-\boldsymbol{B}^{-1}\boldsymbol{NX}_N \tag{3-5}$$

将式(3-5)代入目标函数(3-2)，得到

$$\begin{aligned} z&=\boldsymbol{C}_B(\boldsymbol{B}^{-1}\boldsymbol{b}-\boldsymbol{B}^{-1}\boldsymbol{NX}_N)+\boldsymbol{C}_N\boldsymbol{X}_N\\ &=\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{b}+(\boldsymbol{C}_N-\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{N})\boldsymbol{X}_N \end{aligned} \tag{3-6}$$

令非基变量 $\boldsymbol{X}_N=0$，可得到基可行解 $\boldsymbol{X}^{(1)}=\begin{pmatrix}\boldsymbol{B}^{-1}\boldsymbol{b}\\ 0\end{pmatrix}$；目标函数 $z=\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{b}$

用矩阵描述时，θ 规则的表达式是

$$\theta=\min_i\left[\frac{(\boldsymbol{B}^{-1}\boldsymbol{b})_i}{(\boldsymbol{B}^{-1}\boldsymbol{P}_k)_i}\mid(\boldsymbol{B}^{-1}\boldsymbol{P}_k)_i>0\right]=\frac{(\boldsymbol{B}^{-1}\boldsymbol{b})_l}{(\boldsymbol{B}^{-1}\boldsymbol{P}_k)_l} \tag{3-7}$$

这里的$(\boldsymbol{B}^{-1}\boldsymbol{b})_i$表示$(\boldsymbol{B}^{-1}\boldsymbol{b})$中的第 i 个元素；换入变量 x_k列向量是$(\boldsymbol{B}^{-1}\boldsymbol{P}_k)$，$(\boldsymbol{B}^{-1}\boldsymbol{P}_k)_l$表示向量$(\boldsymbol{B}^{-1}\boldsymbol{P}_k)$中第 l 个元素，其对应的换出变量 x_l。

这里的表达式的形式与第 2 章中有所不同；但其含义完全相同。这里不再重述。

(1) 在初始状态，由于 $\boldsymbol{X}_S$在式(3-2)中的系数是 0，在式(3-6)中隐含基变量 $\boldsymbol{X}_S$的系数$(\boldsymbol{C}_B-\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{B})=0$，若 $\boldsymbol{X}_S$的元素都为基变量时，即 $\boldsymbol{X}_B$检验数为 0。若迭代运算后，$\boldsymbol{X}_S$的部分元素是属于基变量，其对应的系数是 0；属于非基变量的部分元素，其对应的系数是非零。

(2) 从式(3-6)可知非基变量 $\boldsymbol{X}_N$的系数是$(\boldsymbol{C}_N-\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{N})$，非基变量的系数$(\boldsymbol{C}_N-\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{N})$就是第 2 章中用符号 $c_j-z_j(j=1,2,\cdots,n)$表示的检验数。这时有 X_s的某 j 元素属于非基变量时，这时因为 $c_{s_{j\in N}}=0$；$\boldsymbol{P}_{s_{j\in N}}$是松弛变量单位向量，所以对应的系数是$(-\boldsymbol{C}_B\boldsymbol{B}^{-1})_j$。因此所有检验数可以用 $\boldsymbol{C}-\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{A}$ 与$-\boldsymbol{C}_B\boldsymbol{B}^{-1}$表示。

(3) 单纯形表与矩阵表示的关系

经过迭代运算后，只是在构成基变量 $\boldsymbol{X}_B$和非基变量 $\boldsymbol{X}_N$的元素之间进行了换入换出交换和对应的$\boldsymbol{C}_B$，$\boldsymbol{C}_N$元素的相应地交换的改变。但基变量与非基变量的分块关系没有变。所以单纯形表与矩阵表示的关系可用表 3-1 表示。

表 3-1

	基变量 $\boldsymbol{X}_B$	非基变量 $\boldsymbol{X}_N$		等式右边 RHS
系数矩阵	$\boldsymbol{B}^{-1}\boldsymbol{B}=\boldsymbol{I}$	$\boldsymbol{B}^{-1}\boldsymbol{N}$	$\boldsymbol{B}^{-1}\boldsymbol{P}_{s_{j\in N}}$	$\boldsymbol{B}^{-1}\boldsymbol{b}$
检验数	$\boldsymbol{C}_B-\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{B}=0$	$\boldsymbol{C}_N-\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{N}$	$(-\boldsymbol{C}_B\boldsymbol{B}^{-1})_{j\in N}$	$-\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{b}$

表 3-1 即为迭代后的单纯形计算表。各部分的数字都用 $\boldsymbol{B}^{-1}$来计算的。此外还可以见到，在初始单位矩阵的位置，经过迭代运算后，就是 $\boldsymbol{B}^{-1}$的位置。

3.2 单纯形法的矩阵计算

改进单纯形法的基本点是用矩阵计算。当用单纯形表求解线性规划问题时，每步迭代都要计算表中所有的数字。实际上由第 2 章 2.4.2 节单纯形法计算步骤的第(2)～(5)步看出，真正用到的数字只有基变量的值、非基变量检验数和换入变量列的系数。即表 3-1 中 $\boldsymbol{B}^{-1}\boldsymbol{b}$，$\boldsymbol{C}_{N_1}-\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{N}_1$ 和 $\boldsymbol{B}^{-1}\boldsymbol{N}_1$ 中换入变量的列，可见最关键是 $\boldsymbol{B}^{-1}$矩阵的计算。

下面讲述单纯形法迭代中下一个表中基的逆矩阵 $\boldsymbol{B}_{\text{new}}^{-1}$与本次表中基的逆矩阵 $\boldsymbol{B}^{-1}$的

关系。

设 a_{ij}，a_{lj} 为本次表中的数字，a'_{ij}，a'_{lj} 为迭代后表中数字，a_{lk} 为主元素。由第 2 章 2.3.5 节给出的公式有

$$\begin{cases} a'_{lj}=\dfrac{a_{lj}}{a_{lk}} & (i=l) \\ a'_{ij}=a_{ij}-\dfrac{a_{lj}}{a_{lk}}a_{ik} & (i\neq l) \end{cases}$$

对 $i=1,2,\cdots,m$；$j=1,2,\cdots,n$ 上述关系式可写为

$$\begin{pmatrix} a'_{11} & \cdots & a'_{1m} \\ \vdots & & \vdots \\ a'_{l1} & \cdots & a'_{lm} \\ \vdots & & \vdots \\ a'_{m1} & \cdots & a'_{mm} \end{pmatrix} = \overset{\text{第 } l \text{ 列}}{\begin{pmatrix} 1 & \cdots & -a_{1k}/a_{lk} & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & 1/a_{lk} & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & -a_{mk}/a_{lk} & \cdots & 1 \end{pmatrix}} \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{l1} & \cdots & a_{lm} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mm} \end{pmatrix}$$

因此只要将本次表中换入变量的列 $\begin{pmatrix} a_{1k} \\ \vdots \\ a_{lk} \\ \vdots \\ a_{mk} \end{pmatrix}$ 变换为 $\boldsymbol{\xi}=\begin{pmatrix} -a_{1k}/a_{lk} \\ \vdots \\ 1/a_{lk} \\ \vdots \\ -a_{mk}/a_{lk} \end{pmatrix}$ 将其取代 $m\times m$ 的单位矩阵的第 l 列，再左乘本次表中基的逆矩阵 $\boldsymbol{B}^{-1}$，即可求得下表中新基的逆矩阵 $\boldsymbol{B}_{\text{new}}^{-1}$

以下用例子说明 $\boldsymbol{B}_{\text{new}}^{-1}$ 的求取和单纯形法用矩阵计算的具体计算过程。

例 1 用改进单纯形法，即单纯形法的矩阵计算求解线性规划问题

$$\max z=2x_1+3x_2+0x_3+0x_4+0x_5$$

$$\begin{cases} x_1+2x_2+x_3=8 \\ 4x_1+\qquad x_4=16 \\ 4x_2+x_5=12 \\ x_1,x_2\geqslant 0 \end{cases}$$

解 第一步：利用前节中矩阵描述线性规划问题的表达式，给出初始基

$$\boldsymbol{B}_0=(\boldsymbol{P}_3,\boldsymbol{P}_4,\boldsymbol{P}_5)=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

这是单位矩阵，其逆矩阵也是单位矩阵。初始基变量 $\boldsymbol{X}_{B_0}=\begin{pmatrix} x_3 \\ x_4 \\ x_5 \end{pmatrix}$；对应的系数 $\boldsymbol{C}_{B_0}=$

(0,0,0)；非基变量 $X_{N_0}=\begin{bmatrix}x_1\\x_2\end{bmatrix}$；对应的系数 $C_{N_0}=(2,3)$。计算非基变量的检验数。

$$\sigma_{N_0}=C_{N_0}-C_{B_0}B_0^{-1}N_0=(2,3)-(0,0,0)\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}\begin{pmatrix}1&2\\4&0\\0&4\end{pmatrix}=(2,3)$$

由此可确定 x_2 为换入变量，计算 $\theta=\min\left\{\frac{(B_0^{-1}b)_i}{(B_0^{-1}P_2)_i}\middle|B_0^{-1}P_2>0\right\}=\min\left(\frac{8}{2},-,\frac{12}{4}\right)=3$

对应的换出变量为 x_5，由换入变量 x_2 的系数向量 $P_2=\begin{pmatrix}2\\0\\4\end{pmatrix}$，确定4为主元素，然后计算 $\xi_1=\begin{pmatrix}-1/2\\0\\1/4\end{pmatrix}$，求逆矩阵 $B_1^{-1}=E_1B_0^{-1}=\begin{pmatrix}1&0&-1/2\\0&1&0\\0&0&1/4\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}=\begin{pmatrix}1&0&-1/2\\0&1&0\\0&0&1/4\end{pmatrix}$

计算非基变量(x_1,x_5)的系数矩阵：由 $N_1=\begin{pmatrix}1&0\\4&0\\0&1\end{pmatrix}$；变换为

$$B_1^{-1}N_1=\begin{pmatrix}1&0&-1/2\\0&1&0\\0&0&1/4\end{pmatrix}\begin{pmatrix}1&0\\4&0\\0&1\end{pmatrix}=\begin{pmatrix}1&-1/2\\4&0\\0&1/4\end{pmatrix}$$

并且计算

$$B_1^{-1}b=\begin{pmatrix}1&0&-1/2\\0&1&0\\0&0&1/4\end{pmatrix}\begin{pmatrix}8\\16\\12\end{pmatrix}=\begin{pmatrix}2\\16\\3\end{pmatrix}$$

于是得到新的基 $B_1=(P_3,P_4,P_2)$。

新基变量 $X_{B_1}=\begin{pmatrix}x_3\\x_4\\x_2\end{pmatrix}$，非基变量 $X_{N_1}=\begin{pmatrix}x_1\\x_5\end{pmatrix}$；相应的 $C_{B_1}=(0,0,3)$，$C_{N_1}=(2,0)$。

第二步：计算非基变量的检验数

非基变量 $X_{N_1}=\begin{pmatrix}x_1\\x_5\end{pmatrix}$ 的检验数

$$\sigma_{N_1}=C_{N_1}-C_{B_1}B_1^{-1}N_1=(2,0)-(0,0,3)\begin{pmatrix}1&0&-1/2\\0&1&0\\0&0&1/4\end{pmatrix}\begin{pmatrix}1&0\\4&0\\0&1\end{pmatrix}$$

$$=(2,-3/4)$$

确定对应的换入变量为 x_1，计算

$$\theta=\min\left\{\frac{(\boldsymbol{B}_1^{-1}\boldsymbol{b})_i}{(\boldsymbol{B}_1^{-1}\boldsymbol{P}_1)_i}\middle|\boldsymbol{B}_1^{-1}\boldsymbol{P}_1>0\right\}=\min\left(\frac{2}{1},\frac{16}{4},3\right)=2$$

对应的换出变量为 x_3，由此得到新的基 $\boldsymbol{B}_2=(\boldsymbol{P}_1,\boldsymbol{P}_4,\boldsymbol{P}_2)$。由 x_1 的系数向量 $\boldsymbol{P}_1=\begin{pmatrix}1\\4\\0\end{pmatrix}$，确定以 1 为主元素，计算 $\boldsymbol{\xi}_2=\begin{pmatrix}1\\-4\\0\end{pmatrix}$ 和新的基矩阵的逆矩阵

$$\boldsymbol{B}_2^{-1}=\boldsymbol{E}_2\boldsymbol{B}_1^{-1}=\begin{pmatrix}1&0&0\\-4&1&0\\0&0&1\end{pmatrix}\begin{pmatrix}1&0&-1/2\\0&1&0\\0&0&1/4\end{pmatrix}=\begin{pmatrix}1&0&-1/2\\-4&1&2\\0&0&1/4\end{pmatrix}$$

并且计算

$$\boldsymbol{B}_2^{-1}\boldsymbol{b}=\begin{pmatrix}1&0&-1/2\\-4&1&2\\0&0&1/4\end{pmatrix}\begin{pmatrix}8\\16\\12\end{pmatrix}=\begin{pmatrix}2\\8\\3\end{pmatrix}$$

第三步：计算非基变量(x_3,x_5)的检验数

非基变量的检验数

$$\boldsymbol{\sigma}_{N_2}=\boldsymbol{C}_{N_2}-\boldsymbol{C}_{B_2}\boldsymbol{B}_2^{-1}\boldsymbol{N}_2=(0,0)-(2,0,3)\begin{pmatrix}1&0&-1/2\\-4&1&2\\0&0&1/4\end{pmatrix}\begin{pmatrix}1&0\\0&0\\0&1\end{pmatrix}$$

$$=(-2,1/4)$$

对应的换入变量为 x_5，计算

$$\theta=\min\left\{\frac{(\boldsymbol{B}_2^{-1}\boldsymbol{b})_i}{(\boldsymbol{B}_2^{-1}\boldsymbol{P}_5)_i}\middle|\boldsymbol{B}_2^{-1}\boldsymbol{P}_5>0\right\}=\min\left(-,\frac{8}{2},\frac{3}{1/4}\right)=4$$

对应的换出变量为 x_4，由此得到新的基 $\boldsymbol{B}_3=(\boldsymbol{P}_1,\boldsymbol{P}_5,\boldsymbol{P}_2)$。

这时换入变量 x_5 的系数向量是 $\boldsymbol{B}_2^{-1}\boldsymbol{P}_5=\begin{pmatrix}-1/2\\2\\1/4\end{pmatrix}$，以 2 为主元素，计算 $\boldsymbol{\xi}_3=\begin{pmatrix}1/4\\1/2\\-1/8\end{pmatrix}$。

$\boldsymbol{B}_3$ 的逆矩阵

$$\boldsymbol{B}_3^{-1}=\boldsymbol{E}_3\boldsymbol{B}_2^{-1}=\begin{pmatrix}1&1/4&0\\0&1/2&0\\0&-1/8&1\end{pmatrix}\begin{pmatrix}1&0&-1/2\\-4&1&2\\0&0&1/4\end{pmatrix}=\begin{pmatrix}0&1/4&0\\-2&1/2&1\\1/2&-1/8&0\end{pmatrix}$$

再计算非基变量 $\boldsymbol{X}_{N_3}=(x_3,x_4)$ 的检验数

$$\boldsymbol{\sigma}_{N_3}=\boldsymbol{C}_{N_3}-\boldsymbol{C}_{B_3}\boldsymbol{B}_3^{-1}\boldsymbol{N}_3=(0,0)-(2,0,3)\begin{pmatrix}0&1/4&0\\-2&1/2&1\\1/2&-1/8&0\end{pmatrix}\begin{pmatrix}1&0\\0&1\\0&0\end{pmatrix}$$

$$=(-3/2,-1/8)$$

因检验数都是负值，得到了最优解为

$$X^*=\begin{pmatrix}x_1\\x_5\\x_2\end{pmatrix}=B_3^{-1}b=\begin{pmatrix}0&1/4&0\\-2&1/2&1\\1/2&-1/8&0\end{pmatrix}\begin{pmatrix}8\\16\\12\end{pmatrix}=\begin{pmatrix}4\\4\\2\end{pmatrix}$$

目标函数的值：$z^*=C_BB_3^{-1}b=(2,0,3)\begin{pmatrix}4\\4\\2\end{pmatrix}=14$

将达到最优解时的最优基和对应的系数 C_B 用符号 $z^*=C_{B^*}B^{*-1}b$ 表示。这说明只要找到最优基就可以计算出最优解。

3.3 对偶问题的提出

这里的对偶是指对同一事物(问题)从不同的角度(立场)观察，有两种拟似对立的表述。如“平面中矩形的面积与周长的关系”。可分别表述为：周长一定，面积最大的矩形是正方形；面积一定，周长最短的矩形是正方形。又如同一个数据集的线性规划问题，可以有两种优化的表述：一个企业决策者做生产规划时，可以提出最大利润为目标函数；也可以提出以最小资源消耗为目标。

以第 2 章例 1 为例，这是简化了的生产优化问题。现决策者考虑根据原有的数据集(a,b,c)，如何为有限资源 b_i 的使用建立一套评价指标 y_i(决策变量)，条件是在原生产技术条件下每种产品的利润不小于原规定值 c_j，总目标是资源消耗最小化为最优解。于是可以构造一个模型

$$\min \omega=8y_1+16y_2+12y_3$$
$$\begin{cases}y_1+4y_2\geqslant 2\\2y_1+4y_3\geqslant 3\\y_1,y_2,y_3\geqslant 0\end{cases}\tag{3-8}$$

称这个线性规划问题为例 1 线性规划问题(这里称原问题)的对偶问题。

可见这里的就是对偶变量 y_i；可以取 $y_i>0$ 的任何值，即 y_i的上界为无限大，所以目标函数只存在最小值。

线性规划的对偶概念，在理论上和实际应用上都有重要意义。以下进一步阐述。

3.4 线性规划的对偶理论

以上讨论可直观地了解到原线性规划问题与对偶问题之间的关系；本节将从理论上进一步讨论线性规划的对偶问题。

3.4.1 原问题与对偶问题的关系

对于同一数据集(a,c,b)，可以用“≤”不等式约束条件和目标函数最大化的原问题与

"≥"不等式约束条件和目标函数最小化的对偶问题，建立标准的对应关系。

展开形式的原问题

$$\max z = c_1x_1 + c_2x_2 + \cdots + c_nx_n$$

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \leqslant \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix} \tag{3-9}$$

$$x_1, x_2, \cdots, x_n \geqslant 0$$

对偶问题

$$\min \omega = y_1b_1 + y_2b_2 + \cdots + y_mb_m$$

$$(y_1, y_2, \cdots, y_m) \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \geqslant (c_1, c_2, \cdots, c_n) \tag{3-10}$$

$$y_1, y_2, \cdots, y_m \geqslant 0$$

以上是原问题与对偶问题的标准形式，它们之间的关系可以用表 3-2 表示。

表 3-2

y_i \ x_j	x_1	x_2	$\cdots$	x_n	原关系	$\min \omega$
y_1	a_{11}	a_{12}	$\cdots$	a_{1n}	$\leqslant$	b_1
y_2	a_{21}	a_{22}	$\cdots$	a_{2n}	$\leqslant$	b_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
y_m	a_{m1}	a_{m2}	$\cdots$	a_{mn}	$\leqslant$	b_m
对偶关系	$\geqslant$	$\geqslant$	$\cdots$	$\geqslant$	$\max z = \min \omega$	
$\max z$	c_1	c_2	$\cdots$	c_n		

表 3-2 是将原问题与对偶问题的关系汇总于一个表中，从正面看是原问题，将它转 90°后看是对偶问题。若将第 2 章的原线性规划的系数列成如表 3-2 的形式，这就是表 3-3。

例 2 根据表 3-3 写出原问题与对偶问题的表达式。

表 3-3

y_i \ x_j	x_1	x_2	b
y_1	1	2	8
y_2	4	0	16
y_3	0	4	12
c	2	3	

解 原问题

$$\max z=2x_1+3x_2 \qquad\qquad \min \omega=8y_1+16y_2+12y_3$$

$$\begin{cases} x_1+2x_2\leqslant 8 \\ 4x_1 \leqslant 16 \\ 4x_2\leqslant 12 \\ x_1,x_2\geqslant 0 \end{cases} \Rightarrow \begin{cases} y_1+4y_2 \geqslant 2 \\ 2y_1 +4y_3\geqslant 3 \\ y_1,y_2,y_3\geqslant 0 \end{cases} \tag{3-11}$$

将上述原问题与对偶问题之间的变换关系称为对称形式。一般线性规划问题中遇到非对称形式时,处理如下。

原问题的约束条件中含有等式约束条件时,按以下步骤处理。

设等式约束条件的线性规划问题

$$\max z=\sum_{j=1}^{n} c_j x_j$$

$$\begin{cases} \sum_{j=1}^{n} a_{ij}x_j=b_i, \quad i=1,2,\cdots,m \\ x_j\geqslant 0, \quad j=1,2,\cdots,n \end{cases} \tag{3-12}$$

第一步:先将等式约束条件分解为两个不等式约束条件。这时上述线性规划问题可表示为

$$\max z=\sum_{j=1}^{n} c_j x_j$$

$$\begin{cases} \sum_{j=1}^{n} a_{ij}x_j\leqslant b_i, \qquad i=1,2,\cdots,m & (3\text{-}13) \\ -\sum_{j=1}^{n} a_{ij}x_j\leqslant -b_i, \quad i=1,2,\cdots,m & (3\text{-}14) \\ x_j\geqslant 0, \quad j=1,2,\cdots,m \end{cases}$$

设y_i'是对应式(3-13)的对偶变量

y_i''是对应式(3-14)的对偶变量。这里 $i=1,2,\cdots,m$。

第二步:按对称形式变换关系可写出它的对偶问题

$$\min \omega=\sum_{i=1}^{m} b_i y_i'+\sum_{i=1}^{m}(-b_i y_i'')$$

$$\begin{cases} \sum_{i=1}^{m} a_{ij}y_i'+\sum_{i=1}^{m}(-a_{ij}y_i'')\geqslant c_j, \quad j=1,2,\cdots,n \\ y_i',y_i''\geqslant 0, \quad i=1,2,\cdots,m \end{cases}$$

将上述规划问题的各式整理后得到

$$\min \omega = \sum_{i=1}^{m} b_i(y_i' - y_i'')$$

$$\sum_{i=1}^{m} a_{ij}(y_i' - y_i'') \geqslant c_j, \quad j=1,2,\cdots,n$$

令 $y_i = y_i' - y_i''$，$y_i', y_i'' \geqslant 0$。由此可见，y_i 不受正、负限制。将 y_i 代入上述规划问题，便得到对偶问题

$$\min \omega = \sum_{i=1}^{m} b_i y_i$$

$$\begin{cases} \sum_{i=1}^{m} a_{ij} y_i \geqslant c_j, & j=1,2,\cdots,n \\ y_i \text{ 为无约束}, & i=1,2,\cdots,m \end{cases}$$

综合上述，线性规划的原问题与对偶问题的关系，其变换形式归纳为表 3-4 中所示的对应关系。

表 3-4

原问题（或对偶问题）	对偶问题（或原问题）
目标函数 max z	目标函数 min ω
变量：n 个；$\geqslant 0$；$\leqslant 0$；无约束	约束条件：n 个；$\geqslant$；$\leqslant$；$=$
约束条件：m 个；$\leqslant$；$\geqslant$；$=$	变量：m 个；$\geqslant 0$；$\leqslant 0$；无约束
约束条件右端项	目标函数变量的系数
目标函数变量的系数	约束条件右端项

例 3 试求下述线性规划原问题的对偶问题

$$\min z = 2x_1 + 3x_2 - 5x_3 + x_4$$

$$\begin{cases} x_1 + x_2 - 3x_3 + x_4 \geqslant 5 & ① \\ 2x_1 + 2x_3 - x_4 \leqslant 4 & ② \\ x_2 + x_3 + x_4 = 6 & ③ \\ x_1 \leqslant 0;\quad x_2, x_3 \geqslant 0;\quad x_4 \text{ 无约束} \end{cases}$$

解 设对应于约束条件①，②，③的对偶变量分别为 y_1, y_2, y_3；则由表 3-4 中原问题和对偶问题的对应关系，可以直接写出上述问题的对偶问题，即

$$\max z'=5y_1+4y_2+6y_3$$

$$\begin{cases} y_1+2y_2 \geqslant 2 \\ y_1 + y_3 \leqslant 3 \\ -3y_1+2y_2+y_3 \leqslant -5 \\ y_1-y_2+y_3=1 \\ y_1 \geqslant 0, y_2 \leqslant 0, y_3 \text{ 无约束} \end{cases}$$

3.4.2 对偶问题的基本性质

(1) 对称性

对偶问题的对偶是原问题。

证 设原问题是

$$\max z=\boldsymbol{CX};\ \boldsymbol{AX} \leqslant \boldsymbol{b};\ \boldsymbol{X} \geqslant 0$$

根据对偶问题的对称变换关系,可以找到它的对偶问题是

$$\min \omega=\boldsymbol{Yb};\ \boldsymbol{YA} \geqslant \boldsymbol{C};\ \boldsymbol{Y} \geqslant 0$$

若将上式两边取负号,又因 $-\min \omega=\max(-\omega)$ 可得到

$$\max(-\omega)=-\boldsymbol{Yb};\ -\boldsymbol{YA} \leqslant -\boldsymbol{C};\ \boldsymbol{Y} \geqslant 0$$

根据对称变换关系,得到上式的对偶问题是

$$\min(-\omega')=-\boldsymbol{CX};\ -\boldsymbol{AX} \geqslant -\boldsymbol{b};\ \boldsymbol{X} \geqslant 0$$

又因

$$\min(-\omega')=-\max \omega'$$

可得

$$\max \omega'=\max z=\boldsymbol{CX};\ \boldsymbol{AX} \leqslant \boldsymbol{b};\ \boldsymbol{X} \geqslant 0$$

这就是原问题。

(2) 弱对偶性

若 $\bar{\boldsymbol{X}}$ 是原问题的可行解,$\bar{\boldsymbol{Y}}$ 是对偶问题的可行解。则存在 $\boldsymbol{C}\bar{\boldsymbol{X}} \leqslant \bar{\boldsymbol{Y}}\boldsymbol{b}$。

证 设原问题是

$$\max z=\boldsymbol{CX};\ \boldsymbol{AX} \leqslant \boldsymbol{b};\ \boldsymbol{X} \geqslant 0$$

因 $\bar{\boldsymbol{X}}$ 是原问题的可行解,所以满足约束条件,即

$$\boldsymbol{A}\bar{\boldsymbol{X}} \leqslant \boldsymbol{b}$$

若 $\bar{\boldsymbol{Y}}$ 是给定的一组值,设它是对偶问题的可行解,将 $\bar{\boldsymbol{Y}}$ 左乘上式,得到

$$\bar{\boldsymbol{Y}}\boldsymbol{A}\bar{\boldsymbol{X}} \leqslant \bar{\boldsymbol{Y}}\boldsymbol{b}$$

原问题的对偶问题是

$$\min \omega=\boldsymbol{Yb};\ \boldsymbol{YA} \geqslant \boldsymbol{C};\ \boldsymbol{Y} \geqslant 0$$

因为 $\bar{\boldsymbol{Y}}$ 是对偶问题的可行解,所以满足

$$\bar{Y}A \geqslant C$$

将 $\bar{X}$ 右乘上式，得到

$$\bar{Y}A\bar{X} \geqslant C\bar{X}$$

于是得到

$$C\bar{X} \leqslant \bar{Y}A\bar{X} \leqslant \bar{Y}b$$

(3) 无界性

若原问题(对偶问题)为无界解，则其对偶问题(原问题)无可行解。

证 由弱对偶性显然得。

注意这个问题的性质不存在逆。当原问题(对偶问题)无可行解时，其对偶问题(原问题)或具有无界解或无可行解。例如下述一对问题两者皆无可行解。

原问题(对偶问题)

$$\min \omega = -x_1 - x_2$$

$$\begin{cases} x_1 - x_2 \geqslant 1 \\ -x_1 + x_2 \geqslant 1 \\ x_1, x_2 \geqslant 0 \end{cases}$$

对偶问题(原问题)

$$\max z = y_1 + y_2$$

$$\begin{cases} y_1 - y_2 \leqslant -1 \\ -y_1 + y_2 \leqslant -1 \\ y_1, y_2 \geqslant 0 \end{cases}$$

(4) 可行解是最优解时的性质

设 $\hat{X}$ 是原问题的可行解，$\hat{Y}$ 是对偶问题的可行解，当 $C\hat{X} = \hat{Y}b$ 时，$\hat{X}, \hat{Y}$ 是最优解。

证 若 $C\hat{X} = \hat{Y}b$，根据性质 2 可知：对偶问题的所有可行解 $\bar{Y}$ 都存在 $\bar{Y}b \geqslant C\hat{X}$，因 $C\hat{X} = \hat{Y}b$，所以 $\bar{Y}b \geqslant \hat{Y}b$。可见 $\hat{Y}$ 是使目标函数取值最小的可行解，因而是最优解。同样可证明：对于原问题的所有可行解 $\bar{X}$，存在

$$C\hat{X} = \hat{Y}b \geqslant C\bar{X}$$

所以 $\hat{X}$ 是最优解。

(5) 对偶定理

若原问题有最优解，那么对偶问题也有最优解；且目标函数值相等。

证 设 $\hat{X}$ 是原问题的最优解，它对应的基矩阵 B 必存在 $C - C_B B^{-1} A \leqslant 0$。即得到 $\hat{Y}A \geqslant C$，其中 $\hat{Y} = C_B B^{-1}$。

这时 $\hat{Y}$ 是对偶问题的可行解，它使

$$\omega = \hat{Y}b = C_B B^{-1} b$$

因原问题的最优解是 $\hat{X}$，使目标函数取值

$$z = C\hat{X} = C_B B^{-1} b$$

由此，得到

$$\hat{Y}b=C_BB^{-1}b=C\hat{X}$$

可见$\hat{Y}$是对偶问题的最优解。

(6) 互补松弛性

若$\hat{X}$,$\hat{Y}$分别是原问题和对偶问题的可行解。那么$\hat{Y}X_S=0$ 和$Y_S\hat{X}=0$,当且仅当$\hat{X}$,$\hat{Y}$为最优解。

证 设原问题和对偶问题的标准型是

原问题	对偶问题
$\max z=CX$	$\min \omega=Yb$
$\begin{cases}AX+X_S=b\\X,X_S\geqslant 0\end{cases}$	$\begin{cases}YA-Y_S=C\\Y,Y_S\geqslant 0\end{cases}$

将原问题目标函数中的系数向量 C 用 $C=YA-Y_S$ 代替后，得到

$$z=(YA-Y_S)X=YAX-Y_SX \tag{3-15}$$

将对偶问题的目标函数中系数列向量，用 $b=AX+X_S$ 代替后，得到

$$\omega=Y(AX+X_S)=YAX+YX_S \tag{3-16}$$

若$Y_S\hat{X}=0$,$\hat{Y}X_S=0$；则$\hat{Y}b=\hat{Y}A\hat{X}=C\hat{X}$，由性质 4 可知$\hat{X}$,$\hat{Y}$是最优解。

又若$\hat{X}$,$\hat{Y}$分别是原问题和对偶问题的最优解，根据性质 5，则有

$$C\hat{X}=\hat{Y}A\hat{X}=\hat{Y}b$$

由式(3-15)，式(3-16)可知，必有$\hat{Y}X_S=0$, $Y_S\hat{X}=0$。

(7) 设原问题是

$$\max z=CX;\ AX+X_S=b;\ X,X_S\geqslant 0$$

它的对偶问题是

$$\min \omega=Yb;\ YA-Y_S=C;\ Y,Y_S\geqslant 0$$

则原问题单纯形表的检验数行对应其对偶问题的一个基解，其对应关系见表 3-5。

表 3-5

原问题 X_B	X_N	X_S
检验数 0	$C_N-C_BB^{-1}N$	$-C_BB^{-1}$
对偶问题 Y_{S_1}	$-Y_{S_2}$	$-Y$

这里 Y_{S_1} 是对应原问题中基变量 X_B 的剩余变量，Y_{S_2} 是对应原问题中非基变量 X_N 的剩余变量。

证 设 $\boldsymbol{B}$ 是原问题的一个可行基，于是 $\boldsymbol{A}=(\boldsymbol{B},\boldsymbol{N})$；原问题可以改写为

$$\max z=\boldsymbol{C}_B\boldsymbol{X}_B+\boldsymbol{C}_N\boldsymbol{X}_N$$

$$\begin{cases}\boldsymbol{B}\boldsymbol{X}_B+\boldsymbol{N}\boldsymbol{X}_N+\boldsymbol{X}_S=\boldsymbol{b}\\ \boldsymbol{X}_B,\boldsymbol{X}_N,\boldsymbol{X}_S\geqslant 0\end{cases}$$

相应地对偶问题可表示为

$$\min \omega=\boldsymbol{Y}\boldsymbol{b}$$

$$\begin{cases}\boldsymbol{Y}\boldsymbol{B}-\boldsymbol{Y}_{S_1}=\boldsymbol{C}_B & (3\text{-}17)\\ \boldsymbol{Y}\boldsymbol{N}-\boldsymbol{Y}_{S_2}=\boldsymbol{C}_N & (3\text{-}18)\\ \boldsymbol{Y},\boldsymbol{Y}_{S_1},\boldsymbol{Y}_{S_2}\geqslant 0\end{cases}$$

这里 $\boldsymbol{Y}_S=(\boldsymbol{Y}_{S_1},\boldsymbol{Y}_{S_2})$。

当求得原问题的一个解

$$\boldsymbol{X}_B=\boldsymbol{B}^{-1}\boldsymbol{b}$$

其相应的检验数为 $\boldsymbol{C}_N-\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{N}$ 与 $-\boldsymbol{C}_B\boldsymbol{B}^{-1}$。现分析这些检验数与对偶问题的解之间的关系：令 $\boldsymbol{Y}=\boldsymbol{C}_B\boldsymbol{B}^{-1}$，将它代入式(3-17)，式(3-18)得

$$\boldsymbol{Y}_{S_1}=0$$

$$-\boldsymbol{Y}_{S_2}=\boldsymbol{C}_N-\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{N}$$

这些对应关系，可以在单纯形法计算表看到，在求解原问题时，隐含着同时也获得对偶变量的值。

例 4 已知线性规划问题

$$\max z=x_1+x_2$$

$$\begin{cases}-x_1+x_2+x_3\leqslant 2\\ -2x_1+x_2-x_3\leqslant 1\\ x_1,x_2,x_3\geqslant 0\end{cases}$$

试用对偶理论证明上述线性规划问题无最优解。

证 首先看到该问题存在可行解，例如 $\boldsymbol{X}=(0,0,0)$；而上述问题的对偶问题为

$$\min \omega=2y_1+y_2$$

$$\begin{cases}-y_1-2y_2\geqslant 1\\ y_1+y_2\geqslant 1\\ y_1-y_2\geqslant 0\\ y_1,y_2\geqslant 0\end{cases}$$

由第一约束条件可知对偶问题无可行解，因原问题有可行解，故无最优解。

例 5 已知线性规划问题

$$\min \omega = 2x_1 + 3x_2 + 5x_3 + 2x_4 + 3x_5$$

$$\begin{cases} x_1 + x_2 + 2x_3 + x_4 + 3x_5 \geqslant 4 \\ 2x_1 - x_2 + 3x_3 + x_4 + x_5 \geqslant 3 \\ x_j \geqslant 0, \quad j = 1, 2, \cdots, 5 \end{cases}$$

已知其对偶问题的最优解为 $y_1^* = 4/5, y_2^* = 3/5$；$z = 5$。试用对偶理论找出原问题的最优解。

解 先写出它的对偶问题

$$\max z = 4y_1 + 3y_2$$

$$\begin{cases} y_1 + 2y_2 \leqslant 2 & ① \\ y_1 - y_2 \leqslant 3 & ② \\ 2y_1 + 3y_2 \leqslant 5 & ③ \\ y_1 + y_2 \leqslant 2 & ④ \\ 3y_1 + y_2 \leqslant 3 & ⑤ \\ y_1, y_2 \geqslant 0 \end{cases}$$

将 y_1^*, y_2^* 的值代入约束条件，得②，③，④式为严格不等式；由互补松弛性得 $x_2^* = x_3^* = x_4^* = 0$。因 $y_1, y_2 \geqslant 0$；原问题的两个约束条件应取等式，故有

$$x_1^* + 3x_5^* = 4$$

$$2x_1^* + x_5^* = 3$$

求解后得到 $x_1^* = 1, x_5^* = 1$；故原问题的最优解为

$$\boldsymbol{X}^* = (1, 0, 0, 0, 1)^{\mathrm{T}}; \quad \omega^* = 5$$

3.5 影子价格

前面讲到，在单纯形法的每步迭代中，目标函数取值 $z = \boldsymbol{C}_B \boldsymbol{B}^{-1} \boldsymbol{b}$，和检验数 $\boldsymbol{C}_N - \boldsymbol{C}_B \boldsymbol{B}^{-1} \boldsymbol{N}$ 中都有乘子 $\boldsymbol{Y} = \boldsymbol{C}_B \boldsymbol{B}^{-1}$，那么 $\boldsymbol{Y}$ 的经济意义是什么？

设 $\boldsymbol{B}$ 是 $\{\max z = \boldsymbol{CX} \mid \boldsymbol{AX} \leqslant \boldsymbol{b}, \boldsymbol{X} \geqslant 0\}$ 的最优基，因为

$$z^* = \boldsymbol{C}_B \boldsymbol{B}^{-1} \boldsymbol{b} = \boldsymbol{Y}^* \boldsymbol{b}$$

由此求偏导数，得

$$\frac{\partial z^*}{\partial \boldsymbol{b}} = \boldsymbol{C}_B \boldsymbol{B}^{-1} = \boldsymbol{Y}^*$$

即拉格朗日乘子在最优化时的值。要认识它的经济意义，要分析 $\boldsymbol{C}_B$，$\boldsymbol{B}^{-1}$ 的量纲。$\boldsymbol{C}_B$ 这里经济意义是利润，用元/件表示，$\boldsymbol{B}^{-1}$ 是最优基的逆矩阵，本例原材料 A 的量纲是小时/件，kg/件；它们是消耗定额；而逆矩阵的量纲是件/小时，件/kg，它们是资源转换成产

品的效率，所以 C_BB^{-1} 可表示为单位资源转换成利润的效率。由此可见，原问题目标函数的量纲不同，其经济意义也不同，相应地 C_BB^{-1} 的经济意义也不同。可以给予不同的名称。

从另一方面看，即使在其他条件不变的情况下，用偏导数表示的在最优解时对偶变量 y_i^* 的取值来看，它是单位资源微小变化所引起的目标函数的最优值变化的比值，即梯度。从对偶问题来看，最优解就是资源的梯度的线性组合。即 $z^* = \sum_{i=1}^{m} y_i^* b_i$，它的量纲与原问题的技术经济指标量纲应当是一致的，它们的经济意义也应是一致的。因此出现影子价格、影子利润、最优计算价格等名词。还有其他的名称，如运输问题中的对偶变量称为位势等。所以对偶变量的物理经济解释具有一定的艺术性。

从一般经济意义来看，当资源转化为最佳经济效益时的转化率高低，它是在目标函数最优值中体现的，可称为对目标函数边际贡献。

以下用第 2 章例 1 来说明，影子价格的经济意义和具有的特点。

1. 影子价格反映资源对目标函数的边际贡献，即资源转换成经济效益的效率

由第 2 章例 1 的最终计算表(见表 2-5)可见，$y_1^*=1.5$，$y_2^*=0.125$，$y_3^*=0$。这说明在其他条件不变的情况下，若设备增加一台时，该厂按最优计划安排生产可多获利 1.5 元；原材料 A 增加 1kg，可多获利 0.125 元；原材料 B 增加 1kg，对获利无影响。从图 3-1 可看到，设备增加一台时，代表该约束条件的直线由①移至①′，相应的最优解由(4,2)变为(4,2.5)，目标函数 $z=2\times4+3\times2.5=15.5$，即比原来的增大 1.5。又若原材料 A 增加 1kg 时，代表该约束方程的直线由②移至②′，相应的最优解从(4,2)变为(4.25,1.875)，目标函数 $z=2\times4.25+3\times1.875=14.125$。比原来的增加 0.125。原材料 B 增加 1kg 时，该约束方程的直线由③移至③′，这时的最优解不变。

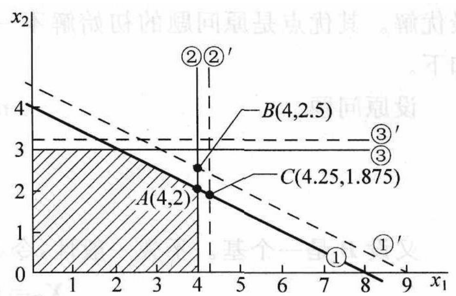


图 3-1

可见资源的影子价格不同，对目标函数值的贡献也不同。从图 3-1 可见，随着资源量的变化，可行域也发生变化；当某资源量的不断增加超过某值时，需要重新计算目标函数的最优解。这将在第 3.7 节灵敏度分析中继续讨论。

2. 影子价格反映了资源的稀缺程度

在不发生退化情况下，$y_1^*=1.5$，$y_2^*=0.125$，$y_3^*=0$，对应资源的剩余量，也即剩余变量的值是：$x_3=0$，$x_4=0$，$x_5=4$。这当某种资源的 $y_i^*>0$，表示这种资源短缺；决策者要增加收入，先注重影子价格高的资源；当某种资源 $y_i^*=0$，表示这种资源有剩余，不短缺，或资源过多供应。可以用影子价格反映在最优解时某种资源利用状态的指标。

3. 影子价格反映了资源的边际使用价值

资源占用者赋予资源的一个内部价格，与资源的市场价格无直接关系。影子价格可

以计算出经济活动的成本,它不是市场价格,不能与资源的市场价格的概念等同。但也有学者在探讨市场价格与影子价格之间的内在关系,这是经济理论问题。

4. 影子价格有三种理论

一是资源最优配置理论;二是机会成本和福利经济学理论;三是全部效益和全部费用理论。影子价格国内外有着不同的论述,这里不作进一步的讨论。

3.6 对偶单纯形法

前节讲到原问题与对偶问题的解之间的对应关系时指出:在单纯形表中进行迭代时,在 b 列中得到的是原问题的基可行解,而在检验数行得到的是对偶问题的基解。通过逐步迭代,当在检验数行得到对偶问题的解也是基可行解时,根据性质(4)、(5)可知,已得到最优解。即原问题与对偶问题都是最优解。

根据对偶问题的对称性,也可以这样考虑:若保持对偶问题的解是基可行解,即 $c_j-\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{P}_j\leqslant 0$,而原问题在非可行解的基础上,通过逐步迭代达到基可行解,这样也得到了最优解。其优点是原问题的初始解不一定是基可行解,可从非基可行解开始迭代,方法如下。

设原问题

$$\max z=\boldsymbol{CX}$$

$$\begin{cases}\boldsymbol{AX}=\boldsymbol{b}\\ \boldsymbol{X}\geqslant 0\end{cases}$$

又设 $\boldsymbol{B}$ 是一个基。不失一般性,令 $\boldsymbol{B}=(\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_m)$,它对应的变量为

$$\boldsymbol{X}_B=(x_1,x_2,\cdots,x_m)$$

当非基变量都为零时,可以得到 $\boldsymbol{X}_B=\boldsymbol{B}^{-1}\boldsymbol{b}$。若在 $\boldsymbol{B}^{-1}\boldsymbol{b}$ 中至少有一个负分量,设 $(\boldsymbol{B}^{-1}\boldsymbol{b})_i<0$,并且在单纯形表的检验数行中的检验数都为非正,即对偶问题保持可行解,它的各分量是

(1) 对应基变量 $x_1,x_2,\cdots,x_m$ 的检验数是

$$\sigma_i=c_i-z_i=c_i-\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{P}_j=0,\quad i=1,2,\cdots,m$$

(2) 对应非基变量 $x_{m+1},\cdots,x_n$ 的检验数是

$$\sigma_j=c_j-z_j=c_j-\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{P}_j\leqslant 0,\quad j=m+1,\cdots,n$$

每次迭代是将基变量中的负分量 x_l 取出,去替换非基变量中的 x_k,经基变换,所有检验数仍保持非正。从原问题来看,经过每次迭代,原问题由非可行解往可行解靠近。当原问题得到可行解时,便得到了最优解。

对偶单纯形法的计算步骤如下:

(1) 对线性规划问题进行变换,使列出的初始单纯形表中所有检验数 $\sigma_j\leqslant 0$,$(j=1,2,\cdots,n)$,即对偶问题为基可行解。

(2) 检查 b 列的数字，若都为非负，检验数都为非正，则已得到最优解。停止计算。若检查 b 列的数字时，至少还有一个负分量，检验数保持非正，那么进行以下计算。

(3) 确定换出变量

按 $\min_i[(\boldsymbol{B}^{-1}\boldsymbol{b})_i \mid (\boldsymbol{B}^{-1}\boldsymbol{b})_i<0]=(\boldsymbol{B}^{-1}\boldsymbol{b})_l$ 对应的基变量 x_l 为换出变量。

(4) 确定换入变量

在单纯形表中检查 x_l 所在行的各系数 $\alpha_{lj}(j=1,2,\cdots,n)$。若所有 $\alpha_{lj}\geqslant 0$，则无可行解，停止计算。若存在 $\alpha_{lj}<0(j=1,2,\cdots,n)$，计算 $\theta=\min_j\left(\frac{c_j-z_j}{\alpha_{lj}}\middle|\alpha_{lj}<0\right)=\frac{c_k-z_k}{\alpha_{lk}}$，按 θ 规则所对应的列的非基变量 x_k 为换入变量，这样才能保持得到的对偶问题解仍为可行解。

(5) 以 α_{lk} 为主元素，按原单纯形法在表中进行迭代运算，得到新的计算表。

重复步骤(2)～(5)。

下面举例来说明具体算法。

例 6　用对偶单纯形法求解

$$\min \omega=2x_1+3x_2+4x_3$$

$$\begin{cases} x_1+2x_2+x_3\geqslant 3\\ 2x_1-x_2+3x_3\geqslant 4\\ x_1,x_2,x_3\geqslant 0\end{cases}$$

解　先将此问题化成下列形式，以便得到对偶问题的初始可行基

$$\max z=-2x_1-3x_2-4x_3$$

$$\begin{cases} -x_1-2x_2-x_3+x_4=-3\\ -2x_1+x_2-3x_3+x_5=-4\\ x_j\geqslant 0,\quad j=1,2,\cdots,5\end{cases}$$

建立此问题的初始单纯形表，见表 3-6。

表　3-6

$c_j\rightarrow$			-2	-3	-4	0	0
$\boldsymbol{C}_B$	$\boldsymbol{X}_B$	$\boldsymbol{b}$	x_1	x_2	x_3	x_4	x_5
0	x_4	-3	-1	-2	-1	1	0
0	x_5	-4	[-2]	1	-3	0	1
c_j-z_j			-2	-3	-4	0	0

从表 3-6 看到，检验数行对应的对偶问题的解是可行解。因 $\boldsymbol{b}$ 列数字为负，故需进行迭代运算。

换出变量的确定：按上述对偶单纯形法计算步骤(3)，计算

$$\min(-3,-4)=-4$$

故 x_5 为换出变量。

换入变量的确定：按上述对偶单纯形法计算步骤(4)，计算

$$\theta=\min\left\{\frac{-2}{-2},-,\frac{-4}{-3}\right\}=\frac{-2}{-2}=1$$

故 x_1 为换入变量。换入、换出变量的所在列、行的交叉处“-2”为主元素。按单纯形法计算步骤进行迭代，得表 3-7。

由表 3-7 看出，对偶问题仍是可行解，而 b 列中仍有负分量。故重复上述迭代步骤，得表 3-8。

表 3-7

$c_j\to$			-2	-3	-4	0	0
C_B	X_B	b	x_1	x_2	x_3	x_4	x_5
0	x_4	-1	0	$[-5/2]$	1/2	1	$-1/2$
-2	x_1	2	1	$-1/2$	3/2	0	$-1/2$
c_j-z_j			0	-4	-1	0	-1

表 3-8

$c_j\to$			-2	-3	-4	0	0
C_B	X_B	b	x_1	x_2	x_3	x_4	x_5
-3	x_2	2/5	0	1	$-1/5$	$-2/5$	1/5
-2	x_1	11/5	1	0	7/5	$-1/5$	$-2/5$
c_j-z_j			0	0	$-9/5$	$-8/5$	$-1/5$

表 3-8 中，b 列数字全为非负，检验数全为非正，故问题的最优解为

$$X^*=(11/5,2/5,0,0,0)^{\mathrm{T}}$$

若对应两个约束条件的对偶变量分别为 y_1 和 y_2，则对偶问题的最优解为

$$Y^*=(y_1^*,y_2^*)=(8/5,1/5)$$

从以上求解过程可以看到对偶单纯形法有以下优点：

(1) 初始解可以是非可行解，当检验数都为负数时，就可以进行基的变换，这时不需要加入人工变量，因此可以简化计算。

(2) 当变量多于约束条件，对这样的线性规划问题，用对偶单纯形法计算可以减少计算工作量，因此对变量较少，而约束条件很多的线性规划问题，可先将它变换成对偶问题，然后用对偶单纯形法求解。

(3) 在灵敏度分析及求解整数规划的割平面法中，有时需要用对偶单纯形法，这样可使问题的处理简化。对偶单纯形法的局限性主要是，对大多数线性规划问题，很难找到一个对偶问题的初始可行基，因而这种方法在求解线性规划问题时很少单独应用。

3.7 灵敏度分析

在以前讨论线性规划问题时，假定 a_{ij}, b_i, c_j 都是常数。但实际上这些系数往往是估计值和预测值。如市场条件变化后，c_j 值就会变化；a_{ij} 往往是因工艺条件的改变而改变；b_i 是根据资源投入后的经济效果决定的一种决策选择。因此提出这样两个问题：当这些系数有一个或几个发生变化时，已求得的线性规划问题的最优解会有什么变化；或者这些系数在什么范围内变化时，线性规划问题的最优解或最优基不变。后一个问题将在 3.8 节参数线性规划中讨论。

显然，当线性规划问题中某一个或某几个系数发生变化后，原来已得结果一般会发生变化。当然可以用单纯形法从头计算，以便得到新的最优解。这样做很麻烦，而且也没有必要。因在单纯形法迭代时，每次运算都和基变量的系数矩阵 $\boldsymbol{B}$ 有关，因此可以把发生变化的个别系数，经过一定计算后直接填入最终计算表中，并进行检查和分析，可按表 3-9 中的几种情况进行处理。

表 3-9

原问题	对偶问题	结论或继续计算的步骤
可行解	可行解	表中的解仍为最优解
可行解	非可行解	用单纯形法继续迭代求最优解
非可行解	可行解	用对偶单纯形法继续迭代求最优解
非可行解	非可行解	引进人工变量，编制新的单纯形表，求最优解

下面就各种情况进行讨论。

3.7.1 资源数量变化的分析

资源数量变化是指资源中某系数 b_r 发生变化，即 $b_r' = b_r + \Delta b_r$。并假设规划问题的其他系数都不变。这样使最终表中原问题的解相应地变化为

$$\boldsymbol{X}_B' = \boldsymbol{B}^{-1}(\boldsymbol{b} + \Delta\boldsymbol{b})$$

这里 $\Delta\boldsymbol{b} = (0, \cdots, \Delta b_r, 0, \cdots, 0)^{\mathrm{T}}$。只要 $\boldsymbol{X}_B' \geqslant 0$，因最终表中检验数不变，故最优基不变，但最优解的值发生了变化，所以 $\boldsymbol{X}_B'$ 为新的最优解。新的最优解的值可允许变化范围用以下方法确定。

$$\boldsymbol{B}^{-1}(\boldsymbol{b} + \Delta\boldsymbol{b}) = \boldsymbol{B}^{-1}\boldsymbol{b} + \boldsymbol{B}^{-1}\Delta\boldsymbol{b}$$

$$= \boldsymbol{B}^{-1}\boldsymbol{b} + \boldsymbol{B}^{-1}\begin{bmatrix} 0 \\ \vdots \\ \Delta b_r \\ \vdots \\ 0 \end{bmatrix}$$

$$B^{-1}\begin{bmatrix}0\\ \vdots\\ \Delta b_r\\ \vdots\\ 0\end{bmatrix}=\begin{bmatrix}\bar{a}_{1r}\Delta b_r\\ \vdots\\ \bar{a}_{ir}\Delta b_r\\ \vdots\\ \bar{a}_{mr}\Delta b_r\end{bmatrix}=\Delta b_r\begin{bmatrix}\bar{a}_{1r}\\ \vdots\\ \bar{a}_{ir}\\ \vdots\\ \bar{a}_{mr}\end{bmatrix}$$

在最终表中求得的经过变化后的 $\boldsymbol{b}$ 列的所有元素，要求 $\bar{b}_i+\bar{a}_{ir}\Delta b_r\geqslant 0, i=1,2,\cdots,m$。由此可得

$$\bar{a}_{ir}\Delta b_r\geqslant -\bar{b}_i,\quad i=1,2,\cdots,m$$

当 $\bar{a}_{ir}>0$ 时，$\Delta b_r\geqslant -\bar{b}_i/\bar{a}_{ir}$；当 $\bar{a}_{ir}<0$ 时，$\Delta b_r\leqslant -\bar{b}_i/\bar{a}_{ir}$；于是得到

$$\max_i\{-\bar{b}_i/\bar{a}_{ir}\mid \bar{a}_{ir}>0\}\leqslant \Delta b_r\leqslant \min_i\{-\bar{b}_i/\bar{a}_{ir}\mid \bar{a}_{ir}<0\}$$

例如求第 2 章例 1 中第二个约束条件 b_2 的变化范围 Δb_2 时，可计算

$$B^{-1}b+B^{-1}\begin{bmatrix}0\\ \Delta b_2\\ 0\end{bmatrix}=\begin{bmatrix}4\\ 4\\ 2\end{bmatrix}+\begin{bmatrix}0.25\\ 0.5\\ -0.125\end{bmatrix}\Delta b_2\geqslant\begin{bmatrix}0\\ 0\\ 0\end{bmatrix}$$

可得 $\Delta b_2\geqslant -4/0.25=-16, \Delta b_2\geqslant -4/0.5=-8, \Delta b_2\leqslant 2/0.125=16$。所以 Δb_2 的变化范围是[−8,16]；显然 b_2 的变化范围是[8,32]。

例 7 从表 2-5 得知第 1 章例 1 中，每设备台时的影子价格为 1.5 元，若该厂又从其他处抽调 4 台时用于生产产品 Ⅰ，Ⅱ。求这时该厂生产产品 Ⅰ，Ⅱ 的最优方案。

解 先计算 $B^{-1}\Delta b$

$$B^{-1}\Delta b=\begin{bmatrix}0 & 0.25 & 0\\ -2 & 0.5 & 1\\ 0.5 & -0.125 & 0\end{bmatrix}\begin{bmatrix}4\\ 0\\ 0\end{bmatrix}=\begin{bmatrix}0\\ -8\\ 2\end{bmatrix}$$

将上述结果反映到最终表 2-5 中，得表 3-10。

表 3-10

$c_j\rightarrow$			2	3	0	0	0
C_B	X_B	b	x_1	x_2	x_3	x_4	x_5
2	x_1	4+0	1	0	0	0.25	0
0	x_5	4−8	0	0	[−2]	0.5	1
3	x_2	2+2	0	1	0.5	−0.125	0
c_j-z_j			0	0	−1.5	−0.125	0

由于表 3-10 中 **b** 列有负数，故用对偶单纯形法求新的最优解。计算结果见表 3-11。

表 3-11

$c_j \to$			2	3	0	0	0
C_B	X_B	b	x_1	x_2	x_3	x_4	x_5
2	x_1	4	1	0	0	0.25	0
0	x_3	2	0	0	1	−0.25	−0.5
3	x_2	3	0	1	0	0	0.25
$c_j - z_j$			0	0	0	−0.5	−0.75

即该厂最优生产方案应改为生产 4 件产品Ⅰ，生产 3 件产品Ⅱ，获利

$$z^* = 4\times 2 + 3\times 3 = 17(\text{元})$$

从表 3-11 看出 $x_3=2$，即设备还有 2 小时未被利用。

3.7.2 目标函数中价值系数 c_j 的变化分析

可以分别就 c_j 是对应的非基变量和基变量两种情况来讨论。

(1) 若 c_j 是非基变量 x_j 的系数，这时它在计算表中所对应的检验数是

$$\sigma_j = c_j - \boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{P}_j$$

或

$$\sigma_j = c_j - \sum_{i=1}^{m} a_{ij} y_i$$

当 c_j 变化 Δc_j 后，要保证最终表中这个检验数仍小于或等于零，即

$$\sigma_j' = c_j + \Delta c_j - \boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{P}_j \leqslant 0$$

那么 $c_j + \Delta c_j \leqslant \boldsymbol{Y}\boldsymbol{P}_j$，即 Δc_j 的值必须小于或等于 $\boldsymbol{Y}\boldsymbol{P}_j - c_j$，才可以满足原最优解条件。这就可以确定 Δc_j 的范围了。

(2) 若 c_r 是基变量 x_r 的系数。因 $c_r \in \boldsymbol{C}_B$，当 c_r 变化 Δc_r 时，就引起 $\boldsymbol{C}_B$ 的变化，这时

$$\begin{aligned}(\boldsymbol{C}_B + \Delta\boldsymbol{C}_B)\boldsymbol{B}^{-1}\boldsymbol{A} &= \boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{A} + (0,\cdots,\Delta c_r,\cdots,0)\boldsymbol{B}^{-1}\boldsymbol{A} \\ &= \boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{A} + \Delta c_r(\alpha_{r1},\alpha_{r2},\cdots,\alpha_{rn})\end{aligned}$$

可见，当 c_r 变化 Δc_r 后，最终表中的检验数是

$$\sigma_j' = c_j - \boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{P}_j - \Delta c_r \bar{\alpha}_{rj}, \quad (j=1,2,\cdots,n)$$

若要求原最优解不变，即必须满足 $\sigma_j' \leqslant 0$。于是得到

当

$$\begin{aligned}&\bar{\alpha}_{rj} < 0, \quad \Delta c_r \leqslant \sigma_j / \bar{\alpha}_{rj}; \\ &\bar{\alpha}_{rj} > 0, \quad \Delta c_r \geqslant \sigma_j / \bar{\alpha}_{rj} \qquad (j=1,2,\cdots,n)\end{aligned}$$

Δc_r 可变化的范围是

$$\max_j\left\{\frac{\sigma_j}{\bar{\alpha}_{rj}} \middle| \bar{\alpha}_{rj} > 0\right\} \leqslant \Delta c_r \leqslant \min_j\left\{\frac{\sigma_j}{\bar{\alpha}_{rj}} \middle| \bar{\alpha}_{rj} < 0\right\}$$

例 8 试以第 2 章例 1 的最终表 2-5 为例。设基变量 x_2 的系数 c_2 变化 Δc_2，在原最优解不变的条件下，确定 Δc_2 的变化范围。

解 这时表 2-5 最终计算表便成为表 3-12 所示。

表 3-12

c_j			2	$3+\Delta c_2$	0	0	0
C_B	X_B	b	x_1	x_2	x_3	x_4	x_5
2	x_1	4	1	0	0	0.25	0
0	x_5	4	0	0	-2	0.5	1
$3+\Delta c_2$	x_2	2	0	1	0.5	-0.125	0
c_j-z_j			0	0	$-1.5-\Delta c_2/2$	$\Delta c_2/8-1/8$	0

若最优解不变，从表 3-12 可见有

$$-1.5-\Delta c_2/2\leqslant 0 \text{ 和 } \Delta c_2/8-1/8\leqslant 0$$

由此可得 $\Delta c_2\geqslant -1.5/0.5$；$\Delta c_2\leqslant 1$。$\Delta c_2$ 的变化范围为

$$-3\leqslant \Delta c_2\leqslant 1$$

即 x_2 的价值系数 c_2 可以在[0,4]之间变化，而不影响原最优解。

3.7.3 技术系数 a_{ij} 的变化

分两种情况来讨论技术系数 a_{ij} 的变化，一是安排一种新产品，即增加一列 P_j；二是原产品的技术系数发生变化。后者有可能变换后，原问题与对偶问题均为非可行解。下面以具体例子来说明。

例 9 分析在原计划中是否应该安排一种新产品。以第 2 章例 1 为例。设该厂除了生产产品Ⅰ，Ⅱ外，现有一种新产品Ⅲ。已知生产产品Ⅲ，每件需消耗原材料 A，B 各为 6kg，3kg，使用设备 2 台时；每件可获利 5 元。问该厂是否应生产该产品和生产多少?

解 分析该问题的步骤是：

(1) 设生产产品Ⅲ为 x_3' 台，其技术系数向量 $\boldsymbol{P}_3'=(2,6,3)^{\mathrm{T}}$，然后计算最终表中对应 x_3' 的检验数

$$\sigma_3'=c_3'-\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{P}_3'=5-(1.5,0.125,0)(2,6,3)^{\mathrm{T}}=1.25>0$$

说明安排生产产品Ⅲ是有利的。

(2) 计算产品Ⅲ在最终表中对应 x_3' 的列向量

$$\boldsymbol{B}^{-1}\boldsymbol{P}_3'=\begin{bmatrix}0 & 0.25 & 0\\ -2 & 0.5 & 1\\ 0.5 & -0.125 & 0\end{bmatrix}\begin{bmatrix}2\\ 6\\ 3\end{bmatrix}=\begin{bmatrix}1.5\\ 2\\ 0.25\end{bmatrix}$$

并将(1),(2)中的计算结果填入最终计算表 2-5,得表 3-13(a)。

表 3-13(a)

c_j			2	3	0	0	0	5
C_B	X_B	b	x_1	x_2	x_3	x_4	x_5	x_3'
2	x_1	4	1	0	0	0.25	0	1.5
0	x_5	4	0	0	−2	0.5	1	[2]
3	x_2	2	0	1	0.5	−0.125	0	0.25
c_j-z_j			0	0	−1.5	−0.125	0	1.25

由于 b 列的数字没有变化,原问题的解是可行解。但检验数行中还有正检验数,说明目标函数值还可以改善。

(3) 将 x_3' 作为换入变量,x_5 作为换出变量,进行迭代,求出最优解。计算结果见表 3-13(b),这时得最优解:$x_1=1,x_2=1.5,x_3'=2$。总的利润为 16.5 元。比原计划增加了 2.5 元。

表 3-13(b)

c_j			2	3	0	0	0	5
C_B	X_B	b	x_1	x_2	x_3	x_4	x_5	x_3'
2	x_1	1	1	0	1.5	−0.125	−0.75	0
5	x_3'	2	0	0	−1	0.25	0.5	1
3	x_2	1.5	0	1	0.75	−0.1875	−0.125	0
c_j-z_j			0	0	−0.25	−0.4375	−0.625	0

例 10 分析原计划生产产品的工艺结构发生变化。仍以第 2 章例 1 为例,若原计划生产产品Ⅰ的工艺结构有了改进,这时有关它的技术系数向量变为 $P_1'=(2,5,2)^T$,每件利润为 4 元,试分析对原最优计划有什么影响?

解 把改进工艺结构的产品Ⅰ看做产品Ⅰ′,设 x_1'为其产量。于是计算在最终表中对应 x_1'的列向量,并以 x_1'代替 x_1。

$$B^{-1}P_1'=\begin{bmatrix}0 & 0.25 & 0\\ -2 & 0.5 & 1\\ 0.5 & -0.125 & 0\end{bmatrix}\begin{bmatrix}2\\5\\2\end{bmatrix}=\begin{bmatrix}1.25\\0.5\\0.375\end{bmatrix}$$

同时计算出 x_1'的检验数为

$$\begin{aligned}c_1'-C_BB^{-1}P_1'&=4-(1.5,0.125,0)(2,5,2)^T\\&=0.375\end{aligned}$$

将以上计算结果填入最终表 x_1 的列向量位置,得表 3-14。

表 3-14

C_B	X_B	b	x_1'	x_2	x_3	x_4	x_5
2	x_1	4	1.25	0	0	0.25	0
0	x_5	4	0.5	0	−2	0.5	1
3	x_2	2	0.375	1	0.5	−0.125	0
c_j-z_j			0.375	0	−1.5	−0.125	0

由表 3-14 可见 x_1'为换入变量，x_1 为换出变量，经过迭代得到表 3-15。

表 3-15

C_B	X_B	b	x_1'	x_2	x_3	x_4	x_5
4	x_1'	3.2	1	0	0	0.2	0
0	x_5	2.4	0	0	−2	0.4	1
3	x_2	0.8	0	1	0.5	−0.2	0
c_j-z_j			0	0	−1.5	−0.2	0

表 3-15 表明原问题和对偶问题的解都是可行解。所以表中的结果已是最优解。即应当生产产品Ⅰ′，3.2 单位；生产产品Ⅱ，0.8 单位。可获利 15.2 元。

注意：若碰到原问题和对偶问题均为非可行解时，就需要引进人工变量后重新求解。

例 11 假设例 10 的产品Ⅰ′的技术系数向量变为 $P_1'=(4,5,2)^{\mathrm{T}}$，而每件获利仍为 4 元。试问该厂应如何安排最优生产方案？

解 方法与例 10 相同，以 x_1'代替 x_1，计算

$$\boldsymbol{B}^{-1}\boldsymbol{P}_1'=\begin{bmatrix}0 & 0.25 & 0\\ -2 & 0.5 & 1\\ 0.5 & -0.125 & 0\end{bmatrix}\begin{bmatrix}4\\5\\2\end{bmatrix}=\begin{bmatrix}1.25\\-3.5\\1.375\end{bmatrix}$$

x_1'的检验数为 $c_1'-\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{P}_1'=4-(1.5,0.125,0)(4,5,2)^{\mathrm{T}}=-2.625$。将这些数字填入最终表 2-5 的 x_1 列的位置，得到表 3-16。

表 3-16

C_B	X_B	b	x_1'	x_2	x_3	x_4	x_5
2	x_1	4	1.25	0	0	0.25	0
0	x_5	4	−3.5	0	−2	0.5	1
3	x_2	2	1.375	1	0.5	−0.125	0
c_j-z_j			−2.625	0	−1.5	−0.125	0

将表 3-16 的 x_1'替换基变量中 x_1，得表 3-17。

表 3-17

C_B	X_B	b	x_1'	x_2	x_3	x_4	x_5
4	x_1'	3.2	1	0	0	0.2	0
0	x_5	15.2	0	0	−2	1.2	1
3	x_2	−2.4	0	1	0.5	−0.4	0
c_j-z_j			0	0	−1.5	0.4	0

从表 3-17 可见原问题和对偶问题都是非可行解。于是引入人工变量 x_6。因在表 3-17中 x_2 所在行，用方程表示时为

$$0x_1'+x_2+0.5x_3-0.4x_4+0x_5=-2.4$$

引入人工变量 x_6 后，便为

$$-x_2-0.5x_3+0.4x_4+x_6=2.4$$

将 x_6 作为基变量代替 x_2，填入表 3-17，得到表 3-18。

表 3-18

C_B	X_B	b	x_1'	x_2	x_3	x_4	x_5	x_6
4	x_1'	3.2	1	0	0	0.2	0	0
0	x_5	15.2	0	0	−2	1.2	1	0
$-M$	x_6	2.4	0	−1	−0.5	[0.4]	0	1
c_j-z_j			0	$3-M$	$-0.5M$	$-0.8+0.4M$	0	0

这时可按单纯形法求解。x_4 为换入变量，x_6 为换出变量。经基变换运算后，得到表 3-19的上表。在表 3-19 的上表中，确定 x_2 为换入变量，x_5 为换出变量。经基变换运算后，得到表 3-19 的下表。此表的所有检验数都为非正，已得最优解。最优生产方案为生产产品Ⅰ′，0.667 单位；产品Ⅱ，2.667 单位，可得最大利润 10.67 元。

表 3-19

c_j			4	3	0	0	0	$-M$
C_B	X_B	b	x_1'	x_2	x_3	x_4	x_5	x_6
4	x_1'	2	1	0.5	0.25	0	0	−0.5
0	x_5	8	0	[3]	−0.5	0	1	−3
0	x_4	6	0	−2.5	−1.25	1	0	2.5
c_j-z_j			0	1	−1	0	0	$-M+2$
4	x_1'	0.667	1	0	0.33	0	$-0.33\times\frac{1}{2}$	0
3	x_2	2.667	0	1	−0.167	0	0.33	−1
0	x_4	12.667	0	0	1.667	1	0.83	0
c_j-z_j			0	0	−0.83	0	−0.33	$-M+3$

除以上介绍的几项分析以外，还可以作增减约束条件等分析。留给读者自己考虑。

3.8* 参数线性规划

灵敏度分析时，主要讨论在最优基不变情况下，确定系数 a_{ij}，b_i，c_j 的变化范围。而参数线性规划是研究这些参数中某一参数连续变化时，使最优解发生变化的各临界点的值。即把某一参数作为参变量，而目标函数在某区间内是这个参变量的线性函数，含这个参变量的约束条件是线性等式或不等式。因此仍可用单纯形法和对偶单纯形法分析参数线性规划问题。其步骤是：

(1) 对含有某参变量 t 的参数线性规划问题。先令 $t=0$，用单纯形法求出最优解；

(2) 用灵敏度分析法，将参变量 t 直接反映到最终表中；

(3) 当参变量 t 连续变大或变小时，观察 b 列和检验数行各数字的变化。若在 b 列首先出现某负值时，则以它对应的变量为换出变量；于是用对偶单纯形法迭代一步。若在检验数行首先出现某正值时，则将它对应的变量为换入变量；用单纯形法迭代一步；

(4) 在经迭代一步后得到的新表上，令参变量 t 继续变大或变小，重复步骤(3)，直到 b 列不能再出现负值，检验数行不能再出现正值为止。

3.8.1 参数 c 的变化

例 12 试分析以下参数线性规划问题。当参数 $t\geqslant 0$ 时的最优解变化。

$$\max z(t)=(3+2t)x_1+(5-t)x_2$$

$$\begin{cases} x_1 \leqslant 4 \\ 2x_2\leqslant 12 \\ 3x_1+2x_2\leqslant 18 \\ x_1, x_2 \geqslant 0 \end{cases}$$

解 将此模型化为标准型

$$\max z(t)=(3+2t)x_1+(5-t)x_2$$

$$\begin{cases} x_1+x_3=4 \\ 2x_2+x_4=12 \\ 3x_1+2x_2+x_5=18 \\ x_j\geqslant 0,\quad j=1,2,\cdots,5 \end{cases}$$

令 $t=0$，用单纯形法求解的结果，见表 3-20。

表 3-20

c_j			3	5	0	0	0
C_B	X_B	b	x_1	x_2	x_3	x_4	x_5
0	x_3	2	0	0	1	1/3	−1/3
5	x_2	6	0	1	0	1/2	0
3	x_1	2	1	0	0	−1/3	1/3
c_j-z_j			0	0	0	−3/2	−1

将 c 的变化直接反映到最终表 3-20 中，得表 3-21。

表 3-21

c_j			$3+2t$	$5-t$	0	0	0
C_B	X_B	b	x_1	x_2	x_3	x_4	x_5
0	x_3	2	0	0	1	1/3	−1/3
$5-t$	x_2	6	0	1	0	1/2	0
$3+2t$	x_1	2	1	0	0	−1/3	1/3
c_j-z_j			0	0	0	$-3/2+\frac{7}{6}t$	$-1-\frac{2}{3}t$

检查上表的检验数行，当 t 值在 $0\leqslant t<9/7$ 变化时，$\sigma_4<0$，为最优解 $(2,6,2,0,0)^{\mathrm{T}}$；在 $t=9/7$ 时 $\sigma_4=0$，出现第一临界点。当 $t>9/7$ 时，出现 $\sigma_4>0$，这时 x_4 作为换入变量，x_3 作为换出变量，用单纯形法迭代一步，得表 3-22。

表 3-22

c_j			$3+2t$	$5-t$	0	0	0
C_B	X_B	b	x_1	x_2	x_3	x_4	x_5
0	x_4	6	0	0	3	1	−1
$5-t$	x_2	3	0	1	−3/2	0	1/2
$3+2t$	x_1	4	1	0	1	0	0
c_j-z_j			0	0	$9/2-\frac{7}{2}t$	0	$-5/2+\frac{1}{2}t$

检查上表的检验数行，当 t 值继续增大，在 $9/7<t<5$ 变化时，在 $\sigma_5<0$ 得最优解 $(4,3,0,6,0)^{\mathrm{T}}$。$t=5$，$\sigma_5=0$ 为第二临界点。当 $t>(5/2)/(1/2)=5$，出现 $\sigma_5>0$ 时，这时 x_5 作为换入变量，x_2 作为换出变量，再用单纯形法迭代一步，得表 3-23。

表 3-23

c_j			$3+2t$	$5-t$	0	0	0
C_B	X_B	b	x_1	x_2	x_3	x_4	x_5
0	x_4	12	0	2	0	1	0
0	x_5	6	0	2	-3	0	1
$3+2t$	x_1	4	1	0	1	0	0
c_j-z_j			0	$5-t$	$-3-2t$	0	0

表 3-23 中 t 继续增大时，恒有 $\sigma_2,\sigma_3<0$，故当 $t\geqslant 5$ 时，最优解为 $(4,0,0,12,6)^{\mathrm{T}}$。

3.8.2 参数 b 的变化分析

例 13 分析以下线性规划问题，当 $t\geqslant 0$ 时，其最优解的变化。

$$\max z=x_1+3x_2$$

$$\begin{cases} x_1+\ x_2\leqslant 6-t \\ -x_1+2x_2\leqslant 6+t \\ x_1,x_2\geqslant 0 \end{cases}$$

解 将上述模型化为标准型

$$\max z=x_1+3x_2$$

$$\begin{cases} x_1+\ x_2+x_3 \quad =6-t \\ -x_1+2x_2+\quad x_4=6+t \\ x_1,x_2,x_3,x_4\geqslant 0 \end{cases}$$

令 $t=0$，用单纯形法求解的结果，见表 3-24。

表 3-24

c_j			1	3	0	0
C_B	X_B	b	x_1	x_2	x_3	x_4
1	x_1	2	1	0	2/3	$-1/3$
3	x_2	4	0	1	1/3	1/3
c_j-z_j			0	0	$-5/3$	$-2/3$

计算

$$\boldsymbol{B}^{-1}\Delta\boldsymbol{b}=\begin{bmatrix} 2/3 & -1/3 \\ 1/3 & 1/3 \end{bmatrix}\begin{bmatrix} -t \\ t \end{bmatrix}=\begin{bmatrix} -t \\ 0 \end{bmatrix}$$

将此计算结果反映到最终表 3-24，得表 3-25。

表 3-25

c_j			1	3	0	0
C_B	X_B	b	x_1	x_2	x_3	x_4
1	x_1	$2-t$	1	0	2/3	−1/3
3	x_2	4	0	1	1/3	1/3
c_j-z_j			0	0	−5/3	−2/3

在表 3-25 中进行分析，当 t 增大至 $t\geqslant 2$ 时，则 $b\leqslant 0$。即 $0\leqslant t\leqslant 2$ 时，最优解为 $(2-t,4,0,0)^{\mathrm{T}}$。当 $t>2$ 时，则 $b_1<0$；故将 x_1 作为换出变量，用对偶单纯形法迭代一步，得表 3-26。

表 3-26

c_j			1	3	0	0
C_B	X_B	b	x_1	x_2	x_3	x_4
0	x_4	$-6+3t$	−3	0	−2	1
3	x_2	$6-t$	1	1	1	0
c_j-z_j			−2	0	−3	0

从表 3-26 可见，当 $t>6$ 时，问题无可行解；当 $2\leqslant t\leqslant 6$ 时，问题的最优解为 $(0,6-t,0,-6+3t)^{\mathrm{T}}$。

习题

3.1 用改进单纯形法(矩阵计算法)求解以下线性规划问题。

(1) $\max z=6x_1-2x_2+3x_3$

$$\begin{cases}2x_1-x_2+2x_3\leqslant 2\\ x_1+4x_3\leqslant 4\\ x_1,x_2,x_3\geqslant 0\end{cases}$$

(2) $\min z=2x_1+x_2$

$$\begin{cases}3x_1+x_2=3\\ 4x_1+3x_2\geqslant 6\\ x_1+2x_2\leqslant 3\\ x_1,x_2\geqslant 0\end{cases}$$

3.2 已知某线性规划问题,用单纯形法计算时得到的中间某两步的计算表见表3-27,试将表中空白处数字填上。

表 3-27

	c_j		3	5	4	0	0	0
C_B	X_B	b	x_1	x_2	x_3	x_4	x_5	x_6
	x_2	8/3	2/3	1	0	1/3	0	0
	x_5	14/3	−4/3	0	5	−2/3	1	0
	x_6	20/3	5/3	0	4	−2/3	0	1
	c_j-z_j		−1/3	0	4	−5/3	0	0
				…				
	x_2					15/41	8/41	−10/41
	x_3					−6/41	5/41	4/41
	x_1					−2/41	−12/41	15/41
	c_j-z_j							

3.3 写出下列线性规划问题的对偶问题,再写出对偶问题的对偶,并验证其即为原问题。

(1) $\min z=2x_1+2x_2+4x_3$

$$\begin{cases}2x_1+3x_2+5x_3\geqslant 2\\3x_1+x_2+7x_3\leqslant 3\\x_1+4x_2+6x_3\leqslant 5\\x_1,x_2,x_3\geqslant 0\end{cases}$$

(2) $\max z=x_1+2x_2+3x_3+4x_4$

$$\begin{cases}-x_1+x_2-x_3-3x_4=5\\6x_1+7x_2+3x_3-5x_4\geqslant 8\\12x_1-9x_2-9x_3+9x_4\leqslant 20\\x_1x_2\geqslant 0,x_3\leqslant 0,x_4\ \text{无约束}\end{cases}$$

(3) $\min z=\sum_{i=1}^{m}\sum_{j=1}^{n}c_{ij}x_{ij}$

$$\begin{cases}\sum_{j=1}^{n}x_{ij}=a_i, & i=1,\cdots,m\\\sum_{i=1}^{m}x_{ij}=b_j, & j=1,\cdots,n\\x_{ij}\geqslant 0\end{cases}$$

(4) $\max z=\sum_{j=1}^{n}c_jx_j$

$$\begin{cases}\sum_{j=1}^{n}a_{ij}x_j\leqslant b_i, & i=1,\cdots,m_1\leqslant m\\\sum_{j=1}^{n}a_{ij}x_j=b_i, & i=m_1+1,m_1+2,\cdots,m\\x_j\geqslant 0, & \text{当}\ j=1,\cdots,n_1\leqslant n\\x_j\ \text{无约束}, & \text{当}\ j=n_1+1,\cdots,n\end{cases}$$

3.4 判断下列说法是否正确,为什么?

(1) 如线性规划的原问题存在可行解,则其对偶问题也一定存在可行解;

(2) 如线性规划的对偶问题无可行解,则原问题也一定无可行解;

(3) 如果线性规划的原问题和对偶问题都具有可行解,则该线性规划问题一定具有有限最优解。

3.5　设线性规划问题1是

$$\max z_1=\sum_{j=1}^{n} c_j x_j$$

$$\begin{cases}\sum_{j=1}^{n} a_{ij}x_j \leqslant b_i, & i=1,2,\cdots,m \\ x_j \geqslant 0, & j=1,2,\cdots,n\end{cases}$$

$(y_1^*,\cdots,y_m^*)$是其对偶问题的最优解。

又设线性规划问题2是

$$\max z_2=\sum_{j=1}^{n} c_i x_j$$

$$\begin{cases}\sum_{j=1}^{n} a_{ij}x_j \leqslant b_i+k_i, & i=1,2,\cdots,m \\ x_j \geqslant 0, & j=1,2,\cdots,n\end{cases}$$

其中 k_i 是给定的常数，求证

$$\max z_2 \leqslant \max z_1+\sum_{i=1}^{m} k_i y_i^*$$

3.6　已知线性规划问题 $\max z=\boldsymbol{CX},\boldsymbol{AX}=\boldsymbol{b},\boldsymbol{X}\geqslant 0$，分别说明发生下列情况时，其对偶问题的解的变化：

(a) 问题的第 k 个约束条件乘上常数 $\lambda(\lambda\neq 0)$；

(b) 将第 k 个约束条件乘上常数 $\lambda(\lambda\neq 0)$后加到第 r 个约束条件上；

(c) 目标函数 $\max z=\lambda\boldsymbol{CX}(\lambda\neq 0)$；

(d) 模型中 $\boldsymbol{X}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$ 用 $\boldsymbol{X}'=(3x_1',x_2,\cdots,x_n)^{\mathrm{T}}$ 代换。

3.7　已知线性规划问题

$$\max z=c_1x_1+c_2x_2+c_3x_3$$

$$\begin{bmatrix}a_{11}\\a_{21}\end{bmatrix}x_1+\begin{bmatrix}a_{12}\\a_{22}\end{bmatrix}x_2+\begin{bmatrix}a_{13}\\a_{23}\end{bmatrix}x_3+\begin{bmatrix}1\\0\end{bmatrix}x_4+\begin{bmatrix}0\\1\end{bmatrix}x_5=\begin{bmatrix}b_1\\b_2\end{bmatrix}$$

$$x_j\geqslant 0,\quad j=1,\cdots,5$$

用单纯形法求解，得到最终单纯形表如表3-28所示，要求：

(1) 求 $a_{11},a_{12},a_{13},a_{21},a_{22},a_{23},b_1,b_2$ 的值；

(2) 求 c_1,c_2,c_3 的值。

表　3-28

$\boldsymbol{X}_B$	$\boldsymbol{b}$	x_1	x_2	x_3	x_4	x_5
x_3	3/2	1	0	1	1/2	−1/2
x_2	2	1/2	1	0	−1	2
c_j-z_j		−3	0	0	0	−4

3.8 已知线性规划问题

$$\max z=2x_1+x_2+5x_3+6x_4 \qquad \text{对偶变量}$$

$$\begin{cases}2x_1 \qquad +x_3+\ x_4\leqslant 8 & y_1\\ 2x_1+2x_2+x_3+2x_4\leqslant 12 & y_2\\ x_j\geqslant 0,\quad j=1,\cdots,4\end{cases}$$

其对偶问题的最优解为 $y_1^*=4, y_2^*=1$，试应用对偶问题的性质，求原问题的最优解。

3.9 试用对偶单纯形法求解下列线性规划问题。

(1) $\min z=x_1+x_2$

$$\begin{cases}2x_1+\ x_2\geqslant 4\\ x_1+7x_2\geqslant 7\\ x_1, x_2\geqslant 0\end{cases}$$

(2) $\min z=3x_1+2x_2+x_3+4x_4$

$$\begin{cases}2x_1+4x_2+5x_3+\ x_4\geqslant 0\\ 3x_1-\ x_2+7x_3-2x_4\geqslant 2\\ 5x_1+2x_2+\ x_3+6x_4\geqslant 15\\ x_1, x_2, x_3, x_4\geqslant 0\end{cases}$$

3.10 现有线性规划问题

$$\max z=-5x_1+5x_2+13x_3$$

$$\begin{cases}-x_1+\ x_2+\ 3x_3\leqslant 20 & ①\\ 12x_1+4x_2+10x_3\leqslant 90 & ②\\ x_1, x_2, x_3\geqslant 0\end{cases}$$

先用单纯形法求出最优解，然后分析在下列各种条件下，最优解分别有什么变化？

(1) 约束条件①的右端常数由 20 变为 30；

(2) 约束条件②的右端常数由 90 变为 70；

(3) 目标函数中 x_3 的系数由 13 变为 8；

(4) 约束中 x_1 的系数列向量由 $\begin{pmatrix}-1\\12\end{pmatrix}$ 变为 $\begin{pmatrix}0\\5\end{pmatrix}$；

(5) 增加一个约束条件③$2x_1+3x_2+5x_3\leqslant 50$；

(6) 将原约束条件②改变为 $10x_1+5x_2+10x_3\leqslant 100$。

3.11 已知某工厂计划生产Ⅰ，Ⅱ，Ⅲ三种产品，各产品需要在 A,B,C 设备上加工，有关数据见表 3-29。试回答：

表 3-29

设备代号	Ⅰ	Ⅱ	Ⅲ	设备有效台时/月
A	8	2	10	300
B	10	5	8	400
C	2	13	10	420
单位产品利润/千元	3	2	2.9	

(1) 如何充分发挥设备能力，使生产赢利最大？

(2) 若为了增加产量，可借用其他工厂的设备 B，每月可借用 60 台时，租金为 1.8 万元，问借用 B 设备是否合算？

(3) 若另有两种新产品Ⅳ，Ⅴ，其中Ⅳ需用设备 A－12 台时；B－5 台时；C－10 台时，单位产品赢利 2.1 千元；新产品Ⅴ需用设备 A－4 台时，B－4 台时，C－12 台时，单位产品赢利 1.87 千元。如 A,B,C 设备台时不增加，分别回答这两种新产品投产在经济上是否合算？

(4) 对产品工艺重新进行设计，改进结构。改进后生产每件产品Ⅰ，需用设备 A－9 台时，设备 B－12 台时，设备 C－4 台时，单位产品赢利 4.5 千元，问这对原计划有何影响？

3.12 分析下列参数规划中当 t 变化时最优解的变化情况。

(1) $\max z(t)=(3-6t)x_1+(2-2t)x_2+(5-5t)x_3\ (t\geqslant 0)$

$$\begin{cases} x_1+2x_2+x_3\leqslant 430 \\ 3x_1+2x_3\leqslant 460 \\ x_1+4x_2\leqslant 420 \\ x_1,x_2,x_3\geqslant 0 \end{cases}$$

(2) $\max z(t)=(7+2t)x_1+(12+t)x_2+(10-t)x_3\ (t\geqslant 0)$

$$\begin{cases} x_1+x_2+x_3\leqslant 20 \\ 2x_1+2x_2+x_3\leqslant 30 \\ x_1,x_2,x_3\geqslant 0 \end{cases}$$

(3) $\max z(t)=2x_1+x_2\ (0\leqslant t\leqslant 25)$

$$\begin{cases} x_1\leqslant 10+2t \\ x_1+x_2\leqslant 25-t \\ x_2\leqslant 10+2t \\ x_1,x_2\geqslant 0 \end{cases}$$

(4) $\max z(t)=21x_1+12x_2+18x_3+15x_4\ (0\leqslant t\leqslant 59)$

$$\begin{cases} 6x_1+3x_2+6x_3+3x_4\leqslant 30+t \\ 6x_1-3x_2+12x_3+6x_4\leqslant 78-t \\ 9x_1+3x_2-6x_3+9x_4\leqslant 135-2t \\ x_j\geqslant 0,\quad j=1,2,3,4 \end{cases}$$

CHAPTER 4
第4章 运输问题

物流和供应链管理是当前管理研究的热点和前沿领域。供应链是一个由物流系统和该供应链中的所有单个组织或企业相关活动组成的网络。为满足供应链中顾客需求，需要对商品服务及相关信息，从产地到消费地高效率低成本地流动及储存进行规划、执行和控制。运筹学中对运输模型的研究为达到上述目的提供相应的理论基础。

4.1 运输问题的数学模型

在经济建设中，经常碰到大宗物资调运问题。如煤、钢铁、木材、粮食等物资，在全国有若干生产基地，根据已有的交通网，应如何制订调运方案，将这些物资运到各消费地点，而总运费要最小。这问题可用以下数学语言描述。

已知有 m 个生产地点 $A_i,i=1,2,\cdots,m$。可供应某种物资，其供应量(产量)分别为 $a_i,i=1,2,\cdots,m$，有 n 个销地 $B_j,j=1,2,\cdots,n$，其需要量分别为 $b_j,j=1,2,\cdots,n$，从 A_i 到 B_j 运输单位物资的运价(单价)为 c_{ij}，这些数据可汇总于产销平衡表和单位运价表中，见表 4-1，表 4-2。有时可把这两表合二为一。

表 4-1

产地	销地				产量
	1	2	$\cdots$	n	
1					a_1
2					a_2
$\vdots$					$\vdots$
m					a_m
销量	b_1	b_2	$\cdots$	b_n	

表 4-2

产地	销地			
	1	2	$\cdots$	n
1	c_{11}	c_{12}	$\cdots$	c_{1n}
2	c_{21}	c_{22}	$\cdots$	c_{2n}
$\vdots$	$\vdots$			
m	c_{m1}	c_{m2}	$\cdots$	c_{mn}

若用 x_{ij} 表示从 A_i 到 B_j 的运量，那么在产销平衡的条件下，要求得总运费最小的调运方案，可求解以下数学模型

$$\min z = \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij}$$

$$\begin{cases} \sum_{i=1}^{m} x_{ij} = b_j, \quad j=1,2,\cdots,n & (4\text{-}1) \\ \sum_{j=1}^{n} x_{ij} = a_i, \quad i=1,2,\cdots,m & (4\text{-}2) \\ x_{ij} \geqslant 0 \end{cases}$$

这就是运输问题的数学模型。它包含 $m\times n$ 个变量，$(m+n)$个约束方程。其系数矩阵的结构比较松散，且特殊。

$$\begin{array}{c} \\ u_1 \\ u_2 \\ \vdots \\ u_m \\ v_1 \\ v_2 \\ \vdots \\ v_n \end{array}
\begin{array}{c}
\begin{array}{cccccccccccccc} x_{11} & x_{12} & \cdots & x_{1n} & x_{21} & x_{22} & \cdots & x_{2n} & \cdots & x_{m1} & x_{m2} & \cdots & x_{mn} \end{array} \\
\left[\begin{array}{ccccccccccccc}
1 & 1 & \cdots & 1 & & & & & & & & & \\
 & & & & 1 & 1 & \cdots & 1 & & & & & \\
 & & & & & & & & \ddots & & & & \\
 & & & & & & & & & 1 & 1 & \cdots & 1 \\
1 & & & & 1 & & & & & 1 & & & \\
 & 1 & & & & 1 & & & & & 1 & & \\
 & & \ddots & & & & \ddots & & & & & \ddots & \\
 & & & 1 & & & & 1 & & & & & 1
\end{array}\right]
\end{array}
\begin{array}{l} \\ \left.\begin{array}{c} \\ \\ \\ \\ \end{array}\right\} m\text{行} \\ \left.\begin{array}{c} \\ \\ \\ \\ \end{array}\right\} n\text{行} \end{array}$$

该系数矩阵中对应于变量 x_{ij} 的系数向量 $\boldsymbol{P}_{ij}$，其分量中除第 i 个和第 $m+j$ 个为 1 以外，其余的都为零。即

$$\boldsymbol{P}_{ij} = (0\cdots1\cdots0\cdots1\cdots0)^{\mathrm{T}} = e_i + e_{m+j}$$

对产销平衡的运输问题，由于有以下关系式存在

$$\sum_{j=1}^{n} b_j = \sum_{j=1}^{n}\left(\sum_{i=1}^{m} x_{ij}\right) = \sum_{i=1}^{m}\left(\sum_{j=1}^{n} x_{ij}\right) = \sum_{i=1}^{m} a_i$$

所以模型最多只有 $m+n-1$ 个独立约束方程。即系数矩阵的秩$\leqslant m+n-1$。由于有以上特征，所以求解运输问题时，可用比较简便的计算方法，习惯上称为表上作业法。

对产销平衡的运输问题，一定存在可行解。又因为所有变量有界，因而存在最优解。

4.2 表上作业法

表上作业法是单纯形法在求解运输问题时的一种简化方法，其实质是单纯形法，故也称运输问题单纯形法。但具体计算和术语有所不同。可归纳为：

(1) 找出初始基可行解。即在有$(m\times n)$格的产销平衡表上按一定的规则，给出 $m+n-1$ 个数字，称为数字格。它们就是初始基变量的取值。

(2) 求各非基变量的检验数,即在表上计算空格的检验数,判别是否达到最优解。如已是最优解,则停止计算,否则转到下一步。

(3) 确定换入变量和换出变量,找出新的基可行解。在表上用闭回路法调整。

(4) 重复(2),(3)直到得到最优解为止。

以上运算都可以在表上完成,下面通过例子说明表上作业法的计算步骤。

例 1 某公司经销甲产品。它下设三个加工厂。每日的产量分别是:A_1 为 7 吨,A_2 为 4 吨,A_3 为 9 吨。该公司把这些产品分别运往四个销售点。各销售点每日销量为:B_1 为 3 吨,B_2 为 6 吨,B_3 为 5 吨,B_4 为 6 吨。已知从各工厂到各销售点的单位产品的运价为表 4-3 所示。问该公司应如何调运产品,在满足各销点的需要量的前提下,使总运费为最少。

解 先画出这问题的产销平衡表和单位运价表,见表 4-3,表 4-4。

表 4-3 单位运价表

加工厂	销地			
	B_1	B_2	B_3	B_4
A_1	3	11	3	10
A_2	1	9	2	8
A_3	7	4	10	5

表 4-4 产销平衡表

产地	销地				产量
	B_1	B_2	B_3	B_4	
A_1					7
A_2					4
A_3					9
销量	3	6	5	6	

4.2.1 确定初始基可行解

这与一般线性规划问题不同。产销平衡的运输问题总是存在可行解。因有

$$\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j = d$$

必存在

$$x_{ij} \geqslant 0, \quad i=1,\cdots,m, \quad j=1,\cdots,n$$

这就是基可行解。又因

$$0 \leqslant x_{ij} \leqslant \min(a_j, b_j)$$

故运输问题必存在最优解。

确定初始基可行解的方法很多，一般希望的方法是既简便，又尽可能接近最优解。下面介绍两种方法：最小元素法和伏格尔(Vogel)法。

1. 最小元素法

这方法的基本思想是就近供应，即从单位运价表中最小的运价开始确定供销关系，然后次小。一直到给出初始基可行解为止。以例1进行讨论。

第一步：从表4-3中找出最小运价为1，这表示先将 A_2 的产品供应给 B_1。因 $a_2 > b_1$，A_2 除满足 B_1 的全部需要外，还可多余1吨产品。在表4-4的 (A_2, B_1) 的交叉格处填上3。得表4-5。并将表4-3的 B_1 列运价划去，得表4-6。

第二步：在表4-6未划去的元素中再找出最小运价2，确定 A_2 多余的1吨供应 B_3，相应地划去 A_2 行运价，并得到表4-7，表4-8。

表 4-5

加工厂	销地				产量
	B_1	B_2	B_3	B_4	
A_1					7
A_2	3				4
A_3					9
销量	3	6	5	6	

表 4-6

产地	销地			
	B_1	B_2	B_3	B_4
A_1	3	11	3	10
A_2	1	9	2	8
A_3	7	4	10	5

表 4-7

加工厂	销地				产量
	B_1	B_2	B_3	B_4	
A_1					7
A_2	3		1		4
A_3					9
销量	3	6	5	6	

表 4-8

产地	销地			
	B_1	B_2	B_3	B_4
A_1	3	11	3	10
A_2	1	9	2	8
A_3	7	4	10	5

第三步：在表 4-8 未划去的元素中再找出最小运价 3；这样一步步地进行下去，直到单位运价表上的所有元素划去为止，最后在产销平衡表上得到一个调运方案，见表 4-9。这方案的总运费为 86 元。

表 4-9

产地	销地				产量
	B_1	B_2	B_3	B_4	
A_1			4	3	7
A_2	3		1		4
A_3		6		3	9
销量	3	6	5	6	

用最小元素法给出的初始解是运输问题的基可行解，其理由为：

(1) 用最小元素法给出的初始解，是从单位运价表中逐次地挑选最小元素，并比较产量和销量。当产大于销，划去该元素所在列。当产小于销，划去该元素所在行。然后在未划去的元素中再找最小元素，再确定供应关系。这样在产销平衡表上每填入一个数字，在运价表上就划去一行或一列。表中共有 m 行 n 列，总共可划$(n+m)$条直线。但当表中只剩一个元素时，这时当在产销平衡表上填这个数字时，而在运价表上同时划去一行和一列。此时把单价表上所有元素都划去了，相应地在产销平衡表上填了$(m+n-1)$个数字。即给出了$(m+n-1)$个基变量的值。

(2) 这$(m+n-1)$个基变量对应的系数列向量是线性独立的。

证 若表中确定的第一个基变量为 $x_{i_1j_1}$ 它对应的系数列向量为

$$\boldsymbol{P}_{i_1j_1}=e_{i_1}+e_{m+j_1}$$

因当给定 $x_{i_1j_1}$ 的值后，将划去第 i_1 行或第 j_1 列，即其后的系数列向量中再不出现 e_{i_1} 或 e_{m+j_1}，因而 $\boldsymbol{P}_{i_1j_1}$ 不可能用解中的其他向量的线性组合表示。类似地给出第二个，…，第$(m+n-1)$个。这$(m+n-1)$个向量都不可能用解中的其他向量的线性组合表示。故这$(m+n-1)$个向量是线性独立的。

用最小元素法给出初始解时，有可能在产销平衡表上填入一个数字后，在单位运价表上同时划去一行和一列。这时就出现退化。关于退化时的处理将在 4.2.4 节中讲述。

2. 伏格尔法

最小元素法的缺点是：可能开始时节省一处的费用，但随后在其他处要多花几倍的

运费。伏格尔法考虑到，一产地的产品假如不能按最小运费就近供应，就考虑次小运费，这就有一个差额。差额越大，说明不能按最小运费调运时，运费增加越多。因而对差额最大处，就应当采用最小运费调运。基于此，伏格尔法的步骤是：

第一步：在表4-3中分别计算出各行和各列的最小运费和次最小运费的差额，并填入该表的最右列和最下行，见表4-10。

表 4-10

产　地	销　地				行差额
	B_1	B_2	B_3	B_4	
A_1	3	11	3	10	0
A_2	1	9	2	8	1
A_3	7	4	10	5	1
列差额	2	5	1	3	

第二步：从行或列差额中选出最大者，选择它所在行或列中的最小元素。在表4-10中 B_2 列是最大差额所在列。B_2 列中最小元素为4，可确定 A_3 的产品先供应 B_2 的需要。得表4-11。同时将运价表中的 B_2 列数字划去。如表4-12所示。

表 4-11

产　地	销　地				产量
	B_1	B_2	B_3	B_4	
A_1					7
A_2					4
A_3		6			9
销量	3	6	5	6	

第三步：对表4-12中未划去的元素再分别计算出各行、各列的最小运费和次最小运费的差额，并填入该表的最右列和最下行。重复第一、二步。直到给出初始解为止。用此法给出例1的初始解列于表4-13。

表 4-12

产　地	销　地				行差额
	B_1	B_2	B_3	B_4	
A_1	3	11	3	10	0
A_2	1	9	2	8	1
A_3	7	4	10	5	2
列差额	2		1	3	

表 4-13

产地	销地				产量
	B_1	B_2	B_3	B_4	
A_1			5	2	7
A_2	3			1	4
A_3		6		3	9
销量	3	6	5	6	

由以上可见：伏格尔法同最小元素法除在确定供求关系的原则上不同外，其余步骤相同。伏格尔法给出的初始解比用最小元素法给出的初始解更接近最优解。

本例用伏格尔法给出的初始解就是最优解。

4.2.2 最优解的判别

判别的方法是计算空格(非基变量)的检验数 $c_{ij}-\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{P}_{ij}\,(i,j\in \mathbf{N})$。因运输问题的目标函数是要求实现最小化，故当所有的 $c_{ij}-\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{P}_{ij}\geqslant 0$ 时，为最优解。下面介绍两种求空格检验数的方法。

1. 闭回路法

在给出调运方案的计算表上，如表 4-13，从每一空格出发找一条闭回路。它是以某空格为起点。用水平或垂直线向前划，当碰到一数字格时可以转 90°后，继续前进，直到回到起始空格为止。简单的闭回路如图 4-1 的(a)，(b)，(c)所示。复杂的是(a)，(b)，(c)图形的组合。

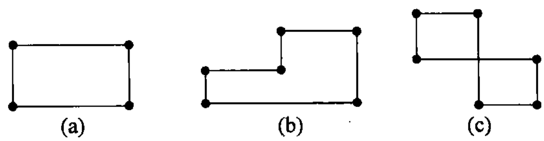


图 4-1

从每一空格出发一定存在和可以找到唯一的闭回路。因$(m+n-1)$个数字格(基变量)对应的系数向量是一个基。任一空格(非基变量)对应的系数向量是这个基的线性组合。如 $\boldsymbol{P}_{ij}, i,j\in \mathbf{N}$ 可表示为

$$\begin{aligned}\boldsymbol{P}_{ij}&=e_i+e_{m+j}\\&=e_i+e_{m+k}-e_{m+k}+e_l-e_l+e_{m+s}-e_{m+s}+e_u-e_u+e_{m+j}\\&=(e_i+e_{m+k})-(e_l+e_{m+k})+(e_l+e_{m+s})-(e_u+e_{m+s})+(e_u+e_{m+j})\\&=\boldsymbol{P}_{ik}-\boldsymbol{P}_{lk}+\boldsymbol{P}_{ls}-\boldsymbol{P}_{us}+\boldsymbol{P}_{uj}\end{aligned}$$

其中 $P_{ik}, P_{lk}, P_{ls}, P_{us}, P_{uj} \in B$。而这些向量构成了闭回路（见图 4-2）。

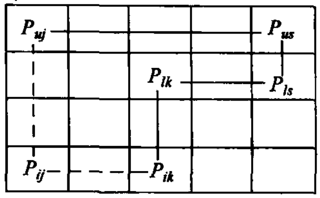


图 4-2

闭回路法计算检验数的经济解释为：在已给出初始解的表 4-9 中，可从任一空格出发，如(A_1,B_1)，若让 A_1 的产品调运 1 吨给 B_1。为了保持产销平衡，就要依次作调整：在(A_1,B_3)处减少 1 吨，(A_2,B_3)处增加 1 吨，(A_2,B_1)处减少 1 吨，即构成了以(A_1,B_1)空格为起点，其他为数字格的闭回路。如表 4-14 中的虚线所示。在这表中闭回路各顶点所在格的右上角数字是单位运价。

表 4-14

产地	销地				产量
	B_1	B_2	B_3	B_4	
A_1	[3] (+1)	[11]	[3] 4(−1)	[10] 3	7
A_2	[1] 3(−1)	[9]	[2] 1(+1)	[8]	4
A_3	[7]	[4] 6	[10]	[5] 3	9
销量	3	6	5	6	

可见这调整的方案使运费增加

$$(+1)\times3+(-1)\times3+(+1)\times2+(-1)\times1=1(\text{元})$$

这表明若这样调整运量将增加运费。将“1”这个数填入(A_1,B_1)格，这就是检验数。按以上所述，可找出所有空格的检验数，见表 4-15。

表 4-15

空格	闭回路	检验数
(11)	(11)—(13)—(23)—(21)—(11)	1
(12)	(12)—(14)—(34)—(32)—(12)	2
(22)	(22)—(23)—(13)—(14)—(34)—(32)—(22)	1
(24)	(24)—(23)—(13)—(14)—(24)	−1
(31)	(31)—(34)—(14)—(13)—(23)—(21)—(31)	10
(33)	(33)—(34)—(14)—(13)—(33)	12

当检验数还存在负数时，说明原方案不是最优解，要继续改进，改进方法见 4.2.3 小节。

2. 位势法

用闭回路法求检验数时，需给每一空格找一条闭回路。当产销点很多时，这种计算很繁。而且在极端情况下，有时找一条闭回路都不容易。下面介绍较为简便的方法——位

势法。

设 $u_1,u_2,\cdots,u_m$; $v_1,v_2,\cdots,v_n$ 是对应产销平衡的运输问题的 $m+n$ 个约束条件的对偶变量。$\boldsymbol{B}$ 是含有一个人工变量 x_a 的 $(m+n)\times(m+n)$ 初始基矩阵。当用最小元素法等给出初始基可行解时,有 $m+n-1$ 个数字格,即赋值基变量;从线性规划的对偶理论可知

$$\boldsymbol{C}_B\boldsymbol{B}^{-1}=(u_1,u_2,\cdots,u_m;\ v_1,v_2,\cdots,v_n)$$

而每个决策变量 x_{ij} 的系数向量 $\boldsymbol{P}_{ij}=e_i+e_{m+j}$,所以 $\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{P}_{ij}=u_i+v_j$。于是检验数

$$\sigma_{ij}=c_{ij}-\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{P}_{ij}=c_{ij}-(u_i+v_j)$$

由单纯形法得知所有基变量的检验数等于0。即

$$c_{ij}-(u_i+v_j)=0,\qquad i,j\in\boldsymbol{B}$$

运输问题的原问题是等式约束;它的对偶问题的变量是无约束。

例如,在例1的由最小元素法得到的初始解中 $x_{23},x_{34},x_{21},x_{32},x_{13},x_{14}$ 是基变量。x_a 为人工变量,这时对应的检验数是

基变量	检验数	
x_{23}	$c_{23}-(u_2+v_3)=0$	即 $2-(u_2+v_3)=0$
x_{34}	$c_{34}-(u_3+v_4)=0$	$5-(u_3+v_4)=0$
x_{21}	$c_{21}-(u_2+v_1)=0$	$1-(u_2+v_1)=0$
x_{32}	$c_{32}-(u_3+v_2)=0$	$4-(u_3+v_2)=0$
x_{13}	$c_{13}-(u_1+v_3)=0$	$3-(u_1+v_3)=0$
x_{14}	$c_{14}-(u_1+v_4)=0$	$10-(u_1+v_4)=0$

以上6个方程,7个未知数为不定方程组,令 $u_1=0$ 可求得

$$u_2=-1,\quad u_3=-5,\quad v_1=2,\quad v_2=9,\quad v_3=3,\quad v_4=10$$

因非基变量的检验数

$$\sigma_{ij}=c_{ij}-(u_i+v_j),\quad i,j\in\mathbf{N}$$

这就可以从已知的 u_i,v_j 值中求得。这些计算可在表格中进行。以例1说明。

第一步:按最小元素法给出表4-9的初始解,并在对应表4-9的数字格处填入单位运价,见表4-16。

表 4-16

产地	销地			
	B_1	B_2	B_3	B_4
A_1			3	10
A_2	1		2	
A_3		4		5

第二步：在表 4-16 上增加一行一列，在列中填入 u_i，在行中填入 v_j，得表 4-17。

表 4-17

产地	销地				u_i
	B_1	B_2	B_3	B_4	
A_1			3	10	0
A_2	1		2		−1
A_3		4		5	−5
v_j	2	9	3	10	

先令 $u_1=0$，然后按 $u_i+v_j=c_{ij}$，$i,j\in \boldsymbol{B}$ 相继地确定 u_i,v_j。分别称 u_i 和 v_j 为行位势和列位势。由表 4-17 可见，当 $u_1=0$ 时，由 $u_1+v_3=3$ 可得 $v_3=3$，由 $u_1+v_4=10$ 可得 $v_4=10$；在 $v_4=10$ 时，由 $u_3+v_4=5$ 可得 $u_3=-5$，以此类推可确定所有的 u_i,v_j 的数值。

第三步：按 $\sigma_{ij}=c_{ij}-(u_i+v_j)$，$i,j\in \mathrm{N}$，计算所有空格的检验数。如

$$\sigma_{11}=c_{11}-(u_1+v_1)=3-(0+2)=1$$

$$\sigma_{12}=c_{12}-(u_1+v_2)=11-(0+9)=2$$

这些计算可直接在表 4-17 上进行。为了方便，特设计计算表，如表 4-18。

表 4-18

产地	销地				u_i
	B_1	B_2	B_3	B_4	
A_1	1 [3]	2 [11]	0 [3]	0 [10]	0
A_2	0 [1]	1 [9]	0 [2]	−1 [8]	−1
A_3	10 [7]	0 [4]	12 [10]	0 [5]	−5
v_j	2	9	3	10	

在表 4-18 中还有负检验数。说明未得最优解，还可以改进。

4.2.3 改进的方法——闭回路调整法

当在表中空格处出现负检验数时，表明未得最优解。若有两个和两个以上的负检验数时，一般选其中最小的负检验数，以它对应的空格为调入格。即以它对应的非基变量为换入变量。由表 4-18 得(2,4)为调入格。以此格为出发点，作一闭回路，如表 4-19 所示。

表 4-19

产地	销地				产量
	B_1	B_2	B_3	B_4	
A_1			4(+1)	3(−1)	7
A_2	3		1(−1)	(+1)	4
A_3		6		3	9
销量	3	6	5	6	

(2,4)格的调入量 θ 是选择闭回路上具有(−1)的数字格中的最小者。即 $\theta=\min(1,3)=1$(其原理与单纯形法中按 θ 规划来确定换出变量相同)。然后按闭回路上的正、负号,加上和减去此值,得到调整方案,如表 4-20 所示。

表 4-20

产地	销地				产量
	B_1	B_2	B_3	B_4	
A_1			5	2	7
A_2	3			1	4
A_3		6		3	9
销量	3	6	5	6	

对表 4-20 给出的解,再用闭回路法或位势法求各空格的检验数,见表 4-21。表中的所有检验数都非负,故表 4-20 中的解为最优解。这时得到的总运费最小是 85 元。

表 4-21

产地	销地			
	B_1	B_2	B_3	B_4
A_1	0	2		
A_2		2	1	
A_3	9		12	

4.2.4 表上作业法计算中的问题

1. 无穷多最优解

在 4.2.1 节中提到,产销平衡的运输问题必定存在最优解。那么有唯一最优解还是

无穷多最优解？判别依据与第 2 章 2.3.3 节讲述的相同。即某个非基变量(空格)的检验数为 0 时，该问题有无穷多最优解。表 4-21 空格(1,1)的检验数是 0，表明例 1 有无穷多最优解。可在表 4-20 中以(1,1)为调入格，作闭回路$(1,1)_+ - (1,4)_- - (2,4)_+ - (2,1)_- - (1,1)_+$。确定$\theta=\min(2,3)=2$。经调整后得到另一最优解，见表 4-22。

表 4-22

产　地	销　地				产量
	B_1	B_2	B_3	B_4	
A_1	2		5		7
A_2	1			3	4
A_3		6		3	9
销量	3	6	5	6	

2. 退化

用表上作业法求解运输问题当出现退化时，在相应的格中一定要填一个 0，以表示此格为数字格。有以下两种情况：

(1) 当确定初始解的各供需关系时，若在(i,j)格填入某数字后，出现A_i处的余量等于B_j处的需量。这时在产销平衡表上填一个数，而在单位运价表上相应地要划去一行和一列。为了使在产销平衡表上有$(m+n-1)$个数字格。这时需要添一个“0”。它的位置可在对应同时划去的那行或那列的任一空格处。如表 4-23，表 4-24 所示。因第一次划去第一列，剩下最小元素为 2，其对应的销地B_2，需要量为 6，而对应的产地A_3未分配量也是 6。这时在产销表(3,2)交叉格中填入 6，这时在单位运价表 4-24 中需同时划去B_2列和A_3行。原则上可在表 4-23 的空格(1,2)，(2,2)，(3,3)，(3,4)中任选一格添加一个 0。但为了减少调整次数，可将 0 添加到上述 4 个空格对应最小运价的位置。从表 4-24 可看出最小运价的位置为 6，故将 0 添加到空格(3,4)处。

表 4-23

产　地	销　地				产量
	B_1	B_2	B_3	B_4	
A_1					7
A_2					4
A_3	3	6		0	9
销量	3	6	5	6	

表 4-24

产地	销地			
	B_1	B_2	B_3	B_4
A_1	3	11	4	5
A_2	7	7	3	8
A_3	1	2	10	6

(2) 在用闭回路法调整时，在闭回路上出现两个和两个以上的具有(−1)标记的相等的最小值。这时经调整后，得到退化解。处理方法为标记最小值处，除有一个变为空格外，在其他最小值的数字格必须填入 0，表明它是基变量。当出现退化解后，并作改进调整时，可能在某闭回路上有标记为(−1)的取值为 0 的数字格，这时应取调整量 $\theta=0$。

4.3 产销不平衡的运输问题及其求解方法

前面讲的表上作业法，都是以产销平衡，即

$$\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j$$

为前提的，但是实际问题中产销往往是不平衡的，就需要把产销不平衡的问题化成产销平衡的问题。

当产大于销

$$\sum_{i=1}^{m} a_i > \sum_{j=1}^{n} b_j$$

时，运输问题的数学模型可写成

$$\min z = \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij}$$

满足

$$\begin{cases} \sum_{j=1}^{n} x_{ij} \leqslant a_i, & (i=1,2,\cdots,m) \\ \sum_{i=1}^{m} x_{ij} = b_j, & (j=1,2,\cdots,n) \\ x_{ij} \geqslant 0 \end{cases}$$

由于总的产量大于销量，就要考虑多余的物资在哪一个产地就地储存的问题。设 $x_{i,n+1}$ 是产地 A_i 的储存量，于是有

$$\sum_{j=1}^{n} x_{ij} + x_{i,n+1} = \sum_{j=1}^{n+1} x_{ij} = a_i, \quad (i=1,\cdots,m)$$

$$\sum_{i=1}^{m} x_{ij} = b_j \qquad (j=1,\cdots,n)$$

$$\sum_{i=1}^{m} x_{i,n+1} = \sum_{i=1}^{m} a_i - \sum_{j=1}^{n} b_j = b_{n+1}$$

令

$$c'_{ij} = c_{ij},\quad 当\ i=1,\cdots,m,\quad j=1,\cdots,n\ 时$$

$$c'_{ij} = 0,\quad 当\ i=1,\cdots,m,\quad j=n+1\ 时$$

将其分别代入，得到

$$\begin{aligned}\min z' &= \sum_{i=1}^{m}\sum_{j=1}^{n+1} c'_{ij}x_{ij} = \sum_{i=1}^{m}\sum_{j=1}^{n} c'_{ij}x_{ij} + \sum_{i=1}^{m} c'_{i,n+1}x_{ij} \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij}\end{aligned}$$

满足

$$\begin{cases}\sum_{j=1}^{n+1} x_{ij} = a_i \\ \sum_{i=1}^{m} x_{ij} = b_j \\ x_{ij} \geqslant 0\end{cases}$$

由于这个模型中

$$\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j + b_{n+1} = \sum_{j=1}^{n+1} b_j$$

所以这是一个产销平衡的运输问题。

由此当产大于销时，只要增加一个假想的销地 $j=n+1$（实际上是储存），该销地总需要量为

$$\sum_{i=1}^{m} a_i - \sum_{j=1}^{n} b_j$$

而在单位运价表中从各产地到假想销地的单位运价为 $c'_{i,n+1}=0$，就转化成一个产销平衡的运输问题。类似地，当销大于产时，可以在产销平衡表中增加一个假想的产地 $i=m+1$，该地产量为

$$\sum_{j=1}^{n} b_j - \sum_{i=1}^{m} a_i$$

在单位运价表上令从该假想产地到各销地的运价 $c'_{m+1,j}=0$，同样可以转化为一个产销平衡的运输问题。

例 2　设有三个化肥厂（A,B,C）供应四个地区（Ⅰ，Ⅱ，Ⅲ，Ⅳ）的农用化肥。假定等量的化肥在这些地区使用效果相同。各化肥厂年产量，各地区年需要量及从各化肥厂到各地区运送单位化肥的运价如表 4-25 所示。试求出总的运费最节省的化肥调拨方案。

表 4-25 万吨

化肥厂	需求地区				产量
	Ⅰ	Ⅱ	Ⅲ	Ⅳ	
A	16	13	22	17	50
B	14	13	19	15	60
C	19	20	23	—	50
最低需求	30	70	0	10	
最高需求	50	70	30	不限	

解 这是一个产销不平衡的运输问题，总产量为160万吨，四个地区的最低需求为110万吨，最高需求为无限。根据现有产量，第Ⅳ个地区每年最多能分配到60万吨，这样最高需求为210万吨，大于产量。为了求得平衡，在产销平衡表中增加一个假想的化肥厂*D*，其年产量为50万吨。由于各地区的需要量包含两部分，如地区Ⅰ，其中30万吨是最低需求，故不能由假想化肥厂*D*供给，令相应运价为*M*(任意大正数)，而另一部分20万吨满足或不满足均可以，因此可以由假想化肥厂*D*供给，按前面讲的，令相应运价为0。对凡是需求分两种情况的地区，实际上可按照两个地区看待。这样可以写出这个问题的产销平衡表(表4-26)和单位运价表(表4-27)。

表 4-26 产销平衡表

产地	销地						产量
	Ⅰ′	Ⅰ″	Ⅱ	Ⅲ	Ⅳ′	Ⅳ″	
A							50
B							60
C							50
D							50
销量	30	20	70	30	10	50	

表 4-27 单位运价表

产地	销地					
	Ⅰ′	Ⅰ″	Ⅱ	Ⅲ	Ⅳ′	Ⅳ″
A	16	16	13	22	17	17
B	14	14	13	19	15	15
C	19	19	20	23	*M*	*M*
D	*M*	0	*M*	0	*M*	0

根据表上作业法计算，可以求得这个问题的最优方案如表4-28所示。

表 4-28

产地	销地						产量
	I′	I″	II	III	IV′	IV″	
A			50				50
B			20		10	30	60
C	30	20	0				50
D				30		20	50
销量	30	20	70	30	10	50	

4.4 应用举例

由于在变量个数相等的情况下，表上作业法的计算远比单纯形法简单得多。所以在解决实际问题时，人们常常尽可能把某些线性规划的问题化为运输问题的数学模型。下面介绍几个典型的例子。

例 3 某厂按合同规定须于当年每个季度末分别提供 10，15，25，20 台同一规格的柴油机。已知该厂各季度的生产能力及生产每台柴油机的成本如表 4-29 所示。又如果生产出来的柴油机当季不交货的，每台每积压一个季度需储存、维护等费用 0.15 万元。要求在完成合同的情况下，作出使该厂全年生产（包括储存、维护）费用最小的决策。

表 4-29

季度	生产能力/台	单位成本/万元
I	25	10.8
II	35	11.1
III	30	11.0
IV	10	11.3

解 由于每个季度生产出来的柴油机不一定当季交货，所以设 x_{ij} 为第 i 季度生产的用于第 j 季度交货的柴油机数。

根据合同要求，必须满足

$$\begin{cases} x_{11} = 10 \\ x_{12} + x_{22} = 15 \\ x_{13} + x_{23} + x_{33} = 25 \\ x_{14} + x_{24} + x_{34} + x_{44} = 20 \end{cases}$$

又每季度生产的用于当季和以后各季交货的柴油机数不可能超过该季度的生产能力，故又有

$$\begin{cases} x_{11}+x_{12}+x_{13}+x_{14}\leqslant 25 \\ \qquad x_{22}+x_{23}+x_{24}\leqslant 35 \\ \qquad\qquad x_{33}+x_{34}\leqslant 30 \\ \qquad\qquad\qquad x_{44}\leqslant 10 \end{cases}$$

第 i 季度生产的用于 j 季度交货的每台柴油机的实际成本 c_{ij} 应该是该季度单位成本加上储存、维护等费用。c_{ij} 的具体数值见表 4-30。

表 4-30　c_{ij} 值

i	j			
	Ⅰ	Ⅱ	Ⅲ	Ⅳ
Ⅰ	10.8	10.95	11.10	11.25
Ⅱ		11.10	11.25	11.40
Ⅲ			11.00	11.15
Ⅳ				11.30

设用 a_i 表示该厂第 i 季度的生产能力，b_j 表示第 j 季度的合同供应量，则问题可写成

$$\min z=\sum_{i=1}^{4}\sum_{j=1}^{4}c_{ij}x_{ij}$$

满足

$$\begin{cases} \sum_{j=1}^{4}x_{ij}\leqslant a_i \\ \sum_{i=1}^{4}x_{ij}=b_j \\ x_{ij}\geqslant 0 \end{cases}$$

显然，这是一个产大于销的运输问题模型。注意到这个问题中当 $i>j$ 时，$x_{ij}=0$，所以应令对应的 $c_{ij}=M$，再加上一个假想的需求 D，就可以把这个问题变成产销平衡的运输模型，并写出产销平衡表和单位运价表(合在一起，见表 4-31)。

表　4-31

产地	销地					产量
	Ⅰ	Ⅱ	Ⅲ	Ⅳ	D	
Ⅰ	10.8	10.95	11.10	11.25	0	25
Ⅱ	M	11.10	11.25	11.40	0	35
Ⅲ	M	M	11.00	11.15	0	30
Ⅳ	M	M	M	11.30	0	10
销量	10	15	25	20	30	

经用表上作业法求解，可得多个最优方案，表 4-32 中列出最优方案之一。即第Ⅰ季度生产 25 台，10 台当季交货，15 台Ⅱ季度交货；Ⅱ季度生产 5 台，用于Ⅲ季度交货；Ⅲ季

度生产 30 台，其中 20 台于当季交货，10 台于Ⅳ季度交货。Ⅳ季度生产 10 台，于当季交货。按此方案生产，该厂总的生产(包括储存、维护)的费用为 773 万元。

表 4-32

生产季度	销售季度					产量
	Ⅰ	Ⅱ	Ⅲ	Ⅳ	D	
Ⅰ	10	15	0			25
Ⅱ			5		30	35
Ⅲ			20	10		30
Ⅳ				10		10
销量	10	15	25	20	30	

例 4 某航运公司承担六个港口城市 A、B、C、D、E、F 的四条固定航线的物资运输任务。已知各条航线的起点、终点城市及每天航班数见表 4-33。假定各条航线使用相同型号的船只，又各城市间的航程天数见表 4-34。

表 4-33

航 线	起点城市	终点城市	每天航班数
1	E	D	3
2	B	C	2
3	A	F	1
4	D	B	1

表 4-34

起点	终 点					
	A	B	C	D	E	F
A	0	1	2	14	7	7
B	1	0	3	13	8	8
C	2	3	0	15	5	5
D	14	13	15	0	17	20
E	7	8	5	17	0	3
F	7	8	5	20	3	0

又知每条船只每次装卸货的时间各需 1 天，则该航运公司至少应配备多少条船，才能满足所有航线的运货需求？

解 该公司所需配备船只分两部分。

(1) 载货航程需要的周转船只数。例如航线 1，在港口 E 装货 1 天，$E \to D$ 航程 17 天，在 D 卸货 1 天，总计 19 天。每天 3 航班，故该航线周转船只需 57 条。各条航线周转

所需船只数见表 4-35。以上累计共需周转船只数 91 条。

表 4-35

航线	装货天数	航程天数	卸货天数	小计	航班数	需周转船只数
1	1	17	1	19	3	57
2	1	3	1	5	2	10
3	1	7	1	9	1	9
4	1	13	1	15	1	15

(2) 各港口间调度所需船只数。有些港口每天到达船数多于需要船数，例如港口 D，每天到达 3 条，需求 1 条；而有些港口到达数少于需求数，例如港口 B。各港口每天余缺船只数的计算见表 4-36。

表 4-36

港口城市	每天到达	每天需求	余缺数
A	0	1	−1
B	1	2	−1
C	2	0	2
D	3	1	2
E	0	3	−3
F	1	0	1

为使配备船只数最少，应做到周转的空船数为最少。因此建立以下运输问题，其产销平衡表见表 4-37。

表 4-37

港　口	A	B	E	每天多余船只
C				2
D				2
F				1
每天缺少船只	1	1	3	

单位运价表应为相应各港口之间的船只航程天数，见表 4-38。

表 4-38

港　口	A	B	E
C	2	3	5
D	14	13	17
F	7	8	3

用表上作业法求出空船的最优调度方案见表 4-39。

表 4-39

港　口	A	B	E	每天多余船只
C	1		1	2
D		1	1	2
F			1	1
每天缺少船只	1	1	3	

由表 4-39 知最少需周转的空船数为 $2\times1+13\times1+5\times1+17\times1+3\times1=40$(条)。这样在不考虑维修、储备等情况下，该公司至少应配备 131 条船。

例 5　在本章的例 1 中，如果假定 ①每个工厂生产的产品不一定直接发运到销售点，可以将其中几个产地集中一起运；②运往各销地的产品可以先运给其中几个销地，再转运给其他销地；③除产、销地之外，中间还可以有几个转运站，在产地之间、销地之间或产地与销地间转运。已知各产地、销地、中间转运站及相互之间每吨产品的运价如表 4-40 所示，问在考虑到产销地之间直接运输和非直接运输的各种可能方案的情况下，如何将三个厂每天生产的产品运往销售地，使总的运费最少。

解　从表 4-40 中看出，从 A_1 到 B_2 每吨产品的直接运费为 11 元，如从 A_1 经 A_3 运往 B_2，每吨运价为 $3+4=7$(元)，从 A_1 经 T_2 运往 B_2 只需 $1+5=6$(元)，而从 A_1 到 B_2 运费最少的路径是从 A_1 经 A_2，B_1 到 B_2，每吨产品的运费只需 $1+1+1=3$(元)。可见这个问题中从每个产地到各销地之间的运输方案是很多的。为了把这个问题仍当做一般的运输问题处理，可以这样做：

表 4-40

项　目		产地			中间转运站				销地			
		A_1	A_2	A_3	T_1	T_2	T_3	T_4	B_1	B_2	B_3	B_4
产地	A_1		1	3	2	1	4	3	3	11	3	10
	A_2	1		—	3	5	—	2	1	9	2	8
	A_3	3	—		1	—	2	3	7	4	10	5
中间转运站	T_1	2	3	1		1	3	2	2	8	4	6
	T_2	1	5	—	1		1	1	4	5	2	7
	T_3	4	—	2	3	1		2	1	8	2	4
	T_4	3	2	3	2	1	2		1	—	2	6
销地	B_1	3	1	7	2	4	1	1		1	4	2
	B_2	11	9	4	8	5	8	—	1		2	1
	B_3	3	2	10	4	2	2	2	4	2		3
	B_4	10	8	5	6	7	4	6	2	1	3	

(1) 由于问题中所有产地、中间转运站、销地都可以看做产地，又可看做销地。因此把整个问题当做有 11 个产地和 11 个销地的扩大的运输问题。

(2) 对扩大的运输问题建立单位运价表。方法将表 4-40 中不可能的运输方案的运价用任意大的正数 M 代替。

(3) 所有中间转运站的产量等于销量。由于运费最少时不可能出现一批物资来回倒运的现象，所以每个转运站的转运数不超过 20 吨。可以规定 T_1, T_2, T_3, T_4 的产量和销量均为 20 吨。由于实际的转运量

$$\sum_{j=1}^{n} x_{ij} \leqslant a_i, \qquad \sum_{i=1}^{m} x_{ij} \leqslant b_j$$

可以在每个约束条件中增加一个松弛变量 x_{ii}，x_{ii} 相当于一个虚构的转运站，意义就是自己运给自己。$(20-x_{ii})$ 就是每个转运站的实际转运量，x_{ii} 的对应运价 $c_{ii}=0$。

(4) 扩大的运输问题中原来的产地与销地因为也有转运站的作用，所以同样在原来产量与销量的数字上加 20 吨，即三个厂每天糖果产量改成 27，24，29 吨，销量均为 20 吨；四个销售点的每天销量改为 23，26，25，26 吨，产量均为 20 吨，同时引进 x_{ii} 作为松弛变量。

下面写出扩大运输问题的两者合一的产销平衡表与单位运价表(见表 4-41)，由于这是一个产销平衡的运输问题，所以可以用表上作业法求解(计算略)。

表 4-41

产地	销地											产量
	A_1	A_2	A_3	T_1	T_2	T_3	T_4	B_1	B_2	B_3	B_4	
A_1	0	1	3	2	1	4	3	3	11	3	10	27
A_2	1	0	M	3	5	M	2	1	9	2	8	24
A_3	3	M	0	1	M	2	3	7	4	10	5	29
T_1	2	3	1	0	1	3	2	2	8	4	6	20
T_2	1	5	M	1	0	1	1	4	5	2	7	20
T_3	4	M	2	3	1	0	2	1	8	2	4	20
T_4	3	2	3	2	1	2	0	1	M	2	6	20
B_1	3	1	7	2	4	1	1	0	1	4	2	20
B_2	11	9	4	8	5	8	M	1	0	2	1	20
B_3	3	2	10	4	2	2	2	4	2	0	3	20
B_4	10	8	5	6	7	4	6	2	1	3	0	20
销量	20	20	20	20	20	20	20	23	26	25	26	

习题

4.1 判断表 4-42 和表 4-43 中给出的调运方案能否作为用表上作业法求解时的初始解？为什么？

表 4-42

产地	销地				产量
	1	2	3	4	
1	0	15			15
2			15	10	25
3	5				5
销量	5	15	15	10	

表 4-43

产地	销地					产量
	1	2	3	4	5	
1	150			250		400
2		200	300			500
3			250		50	300
4	90	210				300
5				80	20	100
销量	240	410	550	330	70	

4.2 判断下列说法是否正确，并说明理由

(1) 在运输问题中，只要任意给出一组含$(m+n+1)$个非零的$\{x_{ij}\}$，且满足 $\sum_{j=1}^{n} x_{ij} = a_i$，$\sum_{i=1}^{m} x_{ij} = b_j$，就可以作为一个初始基可行解；

(2) 表上作业法实质上就是求解运输问题的单纯形法；

(3) 如果运输问题单位运价表的某一行（或某一列）元素分别加上一个常数 k，最优调运方案将不发生变化；

(4) 运输问题单位运价表的全部元素乘上一个常数 $k(k>0)$，最优调运方案将不发生变化。

4.3 用表上作业法和伏格尔（Vogel）法求表 4-44 和表 4-45 中给出的运输问题的最优解和近似最优解（表中数字 M 为任意大正数）。

表 4-44

产地	销地					产量
	甲	乙	丙	丁	戊	
1	10	20	5	9	10	5
2	2	10	8	30	6	6
3	1	20	7	10	4	2
4	8	6	3	7	5	9
销量	4	4	6	2	4	

表 4-45

产地	销地					产量
	甲	乙	丙	丁	戊	
1	10	18	29	13	22	100
2	13	M	21	14	16	120
3	0	6	11	3	M	140
4	9	11	23	18	19	80
5	24	28	36	30	34	60
销量	100	120	100	60	80	

4.4 已知运输问题的产销平衡表、单位运价表及最优调运方案分别见表 4-46 和表 4-47,试回答下列问题。

表 4-46 产销平衡表及最优调运方案

产地	销地				产量
	B_1	B_2	B_3	B_4	
A_1		5		10	15
A_2	0	10	15		25
A_3	5				5
销量	5	15	15	10	

表 4-47 单位运价表

产地	销地			
	B_1	B_2	B_3	B_4
A_1	10	1	20	11
A_2	12	7	9	20
A_3	2	14	16	18

(1) 从 $A_2 \rightarrow B_2$ 的单位运价 c_{22} 在什么范围变化时,上述最优调运方案不发生变化?

(2) $A_2 \rightarrow B_4$ 的单位运价 c_{24} 变为何值时,有无穷多最优调运方案。除表 4-46 中方案

外，至少再写出其他两个。

4.5　某百货公司去外地采购 A、B、C、D 四种规格的服装，数量分别为：A——1 500 套，B——2 000 套，C——3 000 套，D——3 500 套。有三个城市可供应上述规格服装，各城市供应数量分别为：Ⅰ——2 500 套，Ⅱ——2 500 套，Ⅲ——5 000 套。由于这些城市的服装质量、运价和销售情况不同，预计售出后的利润(元/套)也不同，详见表 4-48。请帮助该公司确定一个预期赢利最大的采购方案。

表　4-48

	A	B	C	D
Ⅰ	10	5	6	7
Ⅱ	8	2	7	6
Ⅲ	9	3	4	8

4.6　甲、乙、丙三个城市每年需要煤炭分别为：320 万吨、250 万吨、350 万吨，由 A、B 两处煤矿负责供应。已知煤炭年供应量分别为：A——400 万吨，B——450 万吨。由煤矿至各城市的单位运价(万元/万吨)见表 4-49。由于需大于供，经研究平衡决定，甲城市供应量可减少0～30万吨，乙城市需要量应全部满足，丙城市供应量不少于 270 万吨。试求将供应量分配完又使总运费为最低的调运方案。

表　4-49

	甲	乙	丙
A	15	18	22
B	21	25	16

4.7　某造船厂根据合同要从当年起连续三年年末各提供三条规格型号相同的大型客货轮。已知该厂这三年内生产大型客货轮的能力及每艘客货轮成本如表 4-50 所示。

表　4-50

年　度	正常生产时间内可完成的客货轮数	加班生产时间内可完成的客货轮数	正常生产时每艘成本/万元
1	2	3	500
2	4	2	600
3	1	3	550

已知加班生产时，每艘客货轮成本比正常生产时高出 70 万元。又知造出来的客货轮如当年不交货，每艘每积压一年造成积压损失为 40 万元。在签订合同时，该厂已储存了两艘客货轮，而该厂希望在第三年末完成合同后还能储存一艘备用。问该厂应如何安排每年客货轮的生产量，使在满足上述各项要求的情况下，总的生产费用加积压损失为最少？

CHAPTER 5

第5章

线性目标规划

线性规划在实践中得到广泛的应用，但存在两方面的不足：一是不能处理多目标的优化问题；二是其约束条件过于刚性化，不允许约束资源有丝毫超差。目标规划是为了解决上述不足，而创建的一类数学模型。

目标规划的有关概念和数学模型是由美国学者查纳斯(A. Charnes)和库伯(W. W. Cooper)在1961年首次提出。1965年Yuji. Ijiri在处理多目标问题分析各类目标的重要性时，引入了赋予各目标优先因子和加权系数等概念，进一步完善了数学模型。以后U. Jaahelaineu和Sang. Lee改进了求解方法，近几年来又有新的发展。本章仅介绍有优先等级和加权系数的线性目标规划，下文中所提到的目标规划均指线性目标规划。

5.1 目标规划的数学模型

为了具体说明目标规划与线性规划在处理问题的方法上的区别，先通过例子来介绍目标规划的有关概念及数学模型。

例1 某工厂生产Ⅰ，Ⅱ两种产品，已知有关数据见表5-1。试求获利最大的生产方案。

表 5-1

	Ⅰ	Ⅱ	拥有量
原材料/kg	2	1	11
设备生产能力/小时	1	2	10
利润/(元/件)	8	10	

解 这是求获利最大的单目标的规划问题，用 x_1, x_2 分别表示Ⅰ，Ⅱ产品的产量，其线性规划模型表述为

$$\max z = 8x_1 + 10x_2$$

$$\begin{cases} 2x_1 + x_2 \leqslant 11 \\ x_1 + 2x_2 \leqslant 10 \\ x_1, x_2 \geqslant 0 \end{cases}$$

用图解法求得最优决策方案为：$x_1^* = 4, x_2^* = 3, z^* = 62$(元)。

但实际上工厂在作决策时，还要考虑市场等一系列其他条件。

(1) 根据市场信息，产品Ⅰ的销售量有下降的趋势，故考虑产品Ⅰ的产量不大于产品Ⅱ。

(2) 超过计划供应的原材料时，需用高价采购，会使成本大幅度增加。

(3) 应尽可能充分利用设备台时，但不希望加班。

(4) 应尽可能达到并超过计划利润指标 56 元。

这样在考虑产品决策时，便为多目标决策问题。目标规划方法是解这类决策问题的方法之一。下面引入与建立目标规划数学模型有关的概念。

1. 正、负偏差变量 d^+，d^-

设 x_1，x_2 为决策变量，此外，引进正偏差变量 d^+ 表示决策值超过目标值的部分；负偏差变量 d^- 表示决策值未达到目标值的部分。因决策值不可能既超过目标值同时又未达到目标值，即恒有 $d^+ \cdot d^- = 0$。

2. 绝对约束和目标约束

绝对约束是指必须严格满足的等式约束和不等式约束；如线性规划问题的所有约束条件，不能满足这些约束条件的解称为非可行解，所以它们是硬约束。目标约束是目标规划特有的，可把约束右端项看做要追求的目标值。在达到此目标值时允许发生正或负偏差，因此在这些约束中加入正、负偏差变量，它们是软约束。线性规划问题的目标函数，在给定目标值和加入正、负偏差变量后可变换为目标约束。也可根据问题的需要将绝对约束变换为目标约束。如：例 1 的目标函数 $z=8x_1+10x_2$ 可变换为目标约束 $8x_1+10x_2+d_1^- - d_1^+ = 56$。约束条件 $2x_1+x_2 \leqslant 11$ 可变换为目标约束 $2x_1+x_2+d_2^- - d_2^+ = 11$。

3. 优先因子(优先等级)与权系数

一个规划问题常常有若干目标。但决策者在要求达到这些目标时，是有主次或轻重缓急的不同。要求第一位达到的目标赋予优先因子 P_1，次位的目标赋予优先因子 P_2，…，并规定 $P_k \gg P_{k+1}$，$k=1,2,\cdots,K$。表示 P_k 比 P_{k+1} 有更大的优先权。即首先保证 P_1 级目标的实现，这时可不考虑次级目标；而 P_2 级目标是在实现 P_1 级目标的基础上考虑的；以此类推。若要区别具有相同优先因子的两个目标的差别，这时可分别赋予它们不同的权系数 ω_j，这些都由决策者按具体情况而定。

4. 目标规划的目标函数

目标规划的目标函数(准则函数)是按各目标约束的正、负偏差变量和赋予相应的优先因子及权系数而构造的。当每一目标值确定后，决策者的要求是尽可能缩小偏离目标值。因此目标规划的目标函数只能是 $\min z = f(d^+, d^-)$。其基本形式有三种：

(1) 要求恰好达到目标值，即正、负偏差变量都要尽可能地小，这时

$$\min z = f(d^+ + d^-)$$

(2) 要求不超过目标值，即允许达不到目标值，就是正偏差变量要尽可能地小。这时

$$\min z = f(d^+)$$

(3) 要求超过目标值，即超过量不限，但必须是负偏差变量要尽可能地小，这时

$$\min z=f(d^{-})$$

对每一个具体目标规划问题，可根据决策者的要求和赋予各目标的优先因子来构造目标函数，以下用例子说明。

例 2 例 1 的决策者在原材料供应受严格限制的基础上考虑：首先是产品Ⅱ的产量不低于产品Ⅰ的产量；其次是充分利用设备有效台时，不加班；再次是利润额不小于 56 元。求决策方案。

解 按决策者所要求的，分别赋予这三个目标 P_1,P_2,P_3 优先因子。这问题的数学模型是

$$\min z=P_1d_1^{+}+P_2(d_2^{-}+d_2^{+})+P_3d_3^{-}$$

$$\begin{cases}2x_1+x_2\leqslant 11\\ x_1-x_2+d_1^{-}-d_1^{+}=0\\ x_1+2x_2+d_2^{-}-d_2^{+}=10\\ 8x_1+10x_2+d_3^{-}-d_3^{+}=56\\ x_1,x_2,d_i^{-},d_i^{+}\geqslant 0,\quad i=1,2,3\end{cases}$$

目标规划的一般数学模型为

$$\min z=\sum_{l=1}^{L}P_l\sum_{k=1}^{K}(\omega_{lk}^{-}d_k^{-}+\omega_{lk}^{+}d_k^{+}) \tag{5-1}$$

$$\begin{cases}\sum_{j=1}^{n}c_{kj}x_j+d_k^{-}-d_k^{+}=g_k,\quad k=1,\cdots,K & (5\text{-}2)\\ \sum_{j=1}^{n}a_{ij}x_j\leqslant(=,\geqslant)b_i,\quad i=1,\cdots,m & (5\text{-}3)\\ x_j\geqslant 0,\quad j=1,\cdots,n & (5\text{-}4)\\ d_k^{-},d_k^{+}\geqslant 0,\quad k=1,\cdots,K\end{cases}$$

式中，$\omega_{lk}^{-},\omega_{lk}^{+}$为权系数。

建立目标规划的数学模型时，需要确定目标值、优先等级、权系数等，它都具有一定的主观性和模糊性，可以用专家评定法给以量化。

5.2 解目标规划的图解法

对只具有两个决策变量的目标规划的数学模型，可以用图解法来分析求解。用例 2 来说明。

先在平面直角坐标系的第一象限内，画出各约束条件。绝对约束条件的作图与线性规划相同。本例中满足绝对约束的可行域为三角形 OAB。做目标约束时，先令 $d_i^{-},d_i^{+}=0$，

作相应的直线，然后在这直线旁标上 d_i^-, d_i^+，如图 5-1所示。这表明目标约束可以沿 d_i^-, d_i^+ 所示方向平移。下面根据目标函数中的优先因子来分析求解。首先考虑具有 P_1 优先因子的目标的实现，在目标函数中要求实现 $\min d_1^+$，从图中可见，可以满足 $d_1^+=0$。这时 x_1, x_2 只能在三角形 OBC 的边界和其中取值，接着考虑具有 P_2 优先因子的目标的实现。在目标函数中要求实现 $\min(d_2^+ + d_2^-)$，当 $d_2^+, d_2^-=0$ 时，x_1、x_2 可在线段 ED 上取值。最后考虑具有 P_3 优先因子的目标的实现，在目标函数中要求实现 $\min d_3^-$。从图 5-1 中可以判断可以使 $d_3^-=0$，这就使 x_1、x_2 取值范围缩小到线段 GD 上，这就是该目标规划问题的解。可求得 G 的坐标是(2,4)，D 的坐标是(10/3,10/3)，G，D 的凸线性组合都是该目标规划问题的解。

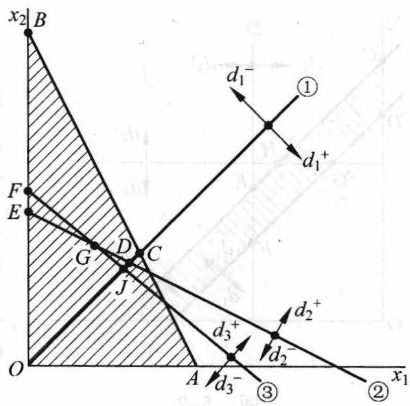


图　5-1

注意目标规划问题求解时，把绝对约束作最高优先级考虑。在本例中能依先后次序都满足 $d_1^+=0, d_2^+ + d_2^-=0, d_3^-=0$，因而 $z^*=0$。但在大多数问题中并非如此，会出现某些约束得不到满足，故将目标规划问题的最优解称为满意解。

例 3　某电视机厂装配黑白和彩色两种电视机，每装配一台电视机需占用装配线 1 小时，装配线每周计划开动 40 小时。预计市场每周彩色电视机的销量是 24 台，每台可获利 80 元；黑白电视机的销量是 30 台，每台可获利 40 元。该厂按预测的销量制定生产计划，其目标为：

第一优先级：充分利用装配线每周计划开动 40 小时；

第二优先级：允许装配线加班；但加班时间每周尽量不超过 10 小时；

第三优先级：装配电视机的数量尽量满足市场需要。因彩色电视机的利润高，取其权系数为 2。

试建立这问题的目标规划模型，并求解黑白和彩色电视机的产量。

解　设 x_1, x_2 分别表示彩色和黑白电视机的产量。

这个问题的目标规划模型为

$$\min z = P_1 d_1^- + P_2 d_2^+ + P_3(2d_3^- + d_4^-)$$

$$\begin{cases} x_1 + x_2 + d_1^- - d_1^+ = 40 \\ x_1 + x_2 + d_2^- - d_2^+ = 50 \\ x_1 \qquad + d_3^- - d_3^+ = 24 \\ \qquad x_2 + d_4^- - d_4^+ = 30 \\ x_1, x_2, d_i^-, d_i^+ \geqslant 0, \quad i=1,\cdots,4 \end{cases}$$

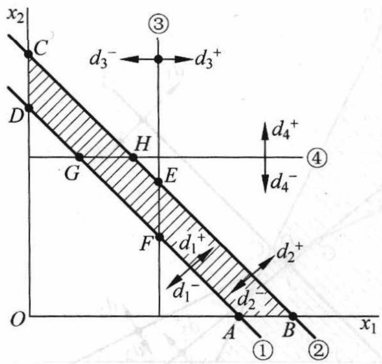


图 5-2

用图解法求解,见图 5-2。

从图 5-2 中看到,在考虑具有 P_1、P_2 的目标实现后,x_1、x_2 的取值范围为 $ABCD$。考虑 P_3 的目标要求为实现 $\min 2d_3^-$;这时 x_1、x_2 取值范围为 $ABEF$;而实现 $\min d_4^-$,x_1、x_2 取值范围为 $CDGH$。因两者无公共区,只能比较最邻近的 H 点和 E 点,看在哪一点处使 $(2d_3^-+d_4^-)$实现最小值。H 点的坐标为(20,30),在该点处 $d_3^-=4$,$d_4^-=0$,有$(2d_3^-+d_4^-)=8$,E 点的坐标为(24,26),在该点处 $d_3^-=0$,$d_4^-=4$,有$(2d_3^-+d_4^-)=4$。

故 E 点为满意解。因其坐标为(24,26),即该厂每周应装配彩色电视机 24 台,黑白电视机 26 台。

5.3 解目标规划的单纯形法

目标规划的数学模型结构与线性规划的数学模型结构形式上没有本质的区别,所以可用单纯形法求解。但要考虑目标规划的数学模型一些特点,作以下规定:

(1) 因目标规划问题的目标函数都是求最小化,所以以 $c_j-z_j\geqslant 0$ $(j=1,2,\cdots,n)$为最优准则。

(2) 因非基变量的检验数中含有不同等级的优先因子,即

$$c_j-z_j=\sum \alpha_{kj}P_k,\quad j=1,2,\cdots,n;\quad k=1,2,\cdots,K$$

因 $P_1\gg P_2\gg\cdots\gg P_K$;从每个检验数的整体来看:检验数的正、负首先决定于 P_1 的系数 α_{1j}的正、负。若 $\alpha_{1j}=0$,这时此检验数的正、负就决定于 P_2 的系数 α_{2j}的正、负,下面可以此类推。

解目标规划问题的单纯形法的计算步骤:

(1) 建立初始单纯形表,在表中将检验数行按优先因子个数分别列成 K 行,置 $k=1$,即对应优先因子行中的第 1 行开始计数。

(2) 检查该行中是否存在负数,且对应列的前 $k-1$ 行的系数是零。若有负数取其中最小者对应的变量为换入变量,转(3)。若无负数,则转(5)。

(3) 按最小比值规则确定换出变量,当存在两个和两个以上相同的最小比值时,选取具有较高优先级别的变量为换出变量。

(4) 按单纯形法进行基变换运算,建立新的计算表,返回(2)。

(5) 当 $k=K$ 时,计算结束。表中的解即为满意解。否则置 $k=k+1$,返回到(2)。

例 4 试用单纯形法来求解例 2。

将例 2 的数学模型化为标准型

$$\min z=P_1d_1^+ + P_2(d_2^- + d_2^+) + P_3d_3^-$$

$$\begin{cases} 2x_1 + x_2 + x_s = 11 \\ x_1 - x_2 + d_1^- - d_1^+ = 0 \\ x_1 + 2x_2 + d_2^- - d_2^+ = 10 \\ 8x_1 + 10x_2 + d_3^- - d_3^+ = 56 \\ x_1, x_2, x_s, d_i^-, d_i^+ \geqslant 0, \quad i=1,2,3 \end{cases}$$

① 取 x_s, d_1^-, d_2^-, d_3^- 为初始基变量，列初始单纯形表，见表 5-2。

② 取 $k=1$，检查 P_1 行的检验数，因该行无负检验数，故转(5)。

表 5-2

c_j							P_1	P_2	P_2	P_3		θ
C_B	X_B	b	x_1	x_2	x_s	d_1^-	d_1^+	d_2^-	d_2^+	d_3^-	d_3^+	
	x_s	11	2	1	1							11/1
	d_1^-	0	1	−1		1	−1					
P_2	d_2^-	10	1	[2]				1	−1			10/2
P_3	d_3^-	56	8	10						1	−1	56/10
c_j-z_j		P_1					1					
		P_2	−1	−2					2			
		P_3	−8	−10							1	

③ 因 $k=1<K=3$，置 $k=k+1=2$，返回到(2)。

④ $k=2$ 时，查出 P_2 行中的检验数有 −1、−2；取 min(−1，−2)＝−2。它对应的变量 x_2 为换入变量，转入(3)。

⑤ 在表 5-2 上计算最小比值

$$\theta=\min(11/1, -, 10/2, 56/10)=10/2$$

它对应的变量 d_2^- 为换出变量，转入(4)。

⑥ 进行基变换运算，计算结果见表 5-3，再返回到(2)。

表 5-3

c_j							P_1	P_2	P_2	P_3		θ
C_B	X_B	b	x_1	x_2	x_s	d_1^-	d_1^+	d_2^-	d_2^+	d_3^-	d_3^+	
	x_s	6	3/2		1			−1/2	1/2			4
P_3	d_1^-	5	3/2			1	−1	1/2	−1/2			10/3
	x_2	5	1/2	1				1/2	−1/2			10
	d_3^-	6	[3]					−5	5	1	−1	6/3
c_j-z_j		P_1					1					
		P_2						1	1			
		P_3	−3					5	−5		1	

⑦ 表 5-3 中 P_1,P_2行检验数全为正,P_3行有两个负值检验数,即变量 x_1 列的 -3 和变量 d_2^+ 列的 -5,因 -5 上面 P_2行的检验数为 1,即 d_2^+ 的检验数应为 $P_2-5P_3>0$,故换入变量只能取 x_1。

⑧ 在表 5-3 中 计算最小比值

$$\theta=\min\left(\frac{6}{3/2},\frac{5}{3/2},\frac{5}{1/2},\frac{6}{3}\right)=\frac{6}{3}$$

其对应变量 d_3^- 为换出变量,转入(4)。

⑨ 进行基变换运算,得新的单纯形表,见表 5-4。

表 5-4

c_j							P_1	P_2	P_2	P_3		θ
C_B	X_B	b	x_1	x_2	x_s	d_1^-	d_1^+	d_2^-	d_2^+	d_3^-	d_3^+	
	x_s	3			1			2	-2	$-1/2$	1/2	6
	d_1^-	2				1	-1	3	-3	$-1/2$	1/2	4
	x_2	4		1				4/3	$-4/3$	$-1/6$	1/6	24
	x_1	2	1					$-5/3$	5/3	1/3	$-1/3$	
c_j-z_j		P_1					1					
		P_2						1	1			
		P_3								1		

表 5-4 所示的解 $x_1^*=2,x_2^*=4$ 为例 2 的满意解。此解相当于图 5-2 的 G 点。检查表 5-4 的检验数行,发现非基变量 d_3^+ 的检验数为 0,这表示存在多重满意解。在表 5-4 中以非基变量 d_3^+ 为换入变量,d_1^- 为换出变量,经迭代得到表 5-5。由表 5-5 得到解 $x_1^*=10/3,x_2^*=10/3$,此解相当于图 5-1 的 D 点,G、D 两点的凸线性组合都是例 2 的满意解。

表 5-5

c_j							P_1	P_2	P_2	P_3		θ
C_B	X_B	b	x_1	x_2	x_s	d_1^-	d_1^+	d_2^-	d_2^+	d_3^-	d_3^+	
	x_s	1			1	-1	1	-1	1			
	d_3^+	4				2	-2	6	-6	-1	1	
	x_2	10/3		1		$-1/3$	1/3	1/3	$-1/3$			
	x_1	10/3	1			2/3	$-2/3$	1/3	$-1/3$			
c_j-z_j		P_1					1					
		P_2						1	1			
		P_3								1		

5.4 应用举例

例 5 某研究所领导在考虑本单位职工的升级调资方案时，依次遵守以下优先级顺序规定：

(1) 不超过年工资总额 3 000 万元；

(2) 提级时，每级的人数不超过定编规定的人数；

(3) Ⅱ，Ⅲ级的升级面尽可能达到现有人数的 20%，且无越级提升；

此外，Ⅲ级不足编制的人数可录用新职工，又Ⅰ级的职工中有 10%要退休。

有关资料汇总于表 5-6 中，问该领导应如何拟订一个满意的方案。

表 5-6

等级	工资额/(万元/年)	现有人数/人	编制人数/人
Ⅰ	10.0	100	120
Ⅱ	7.5	120	150
Ⅲ	5.0	150	150
合计		370	420

解 设 x_1、x_2、x_3 分别表示提升到Ⅰ、Ⅱ级和录用到Ⅲ级的新职工人数。对各目标确定的优先因子为：

P_1——不超过年工资总额 3 000 万元；

P_2——每级的人数不超过定编规定的人数；

P_3——Ⅱ、Ⅲ级的升级面尽可能达到现有人数的 20%。

先分别建立各目标约束。

年工资总额不超过 3 000 万元

$$10(100-100\times0.1+x_1)+7.5(120-x_1+x_2)+5.0(150-x_2+x_3)+d_1^- -d_1^+ =3\,000$$

每级的人数不超过定编规定的人数：

对Ⅰ级有 $100(1-0.1)+x_1+d_2^- -d_2^+ =120$

对Ⅱ级有 $120-x_1+x_2+d_3^- -d_3^+ =150$

对Ⅲ级有 $150-x_2+x_3+d_4^- -d_4^+ =150$

Ⅱ，Ⅲ级的升级面不大于现有人数的 20%，

对Ⅱ级有 $x_1+d_5^- -d_5^+ =120\times0.2$

对Ⅲ级有 $x_2+d_6^- -d_6^+ =150\times0.2$

目标函数：$\min z=P_1d_1^+ +P_2(d_2^+ +d_3^+ +d_4^+)+P_3(d_5^- +d_6^-)$

经过整理后得下列目标规划模型

$$\min z=P_1d_1^+ +P_2(d_2^+ +d_3^+ +d_4^+)+P_3(d_5^- +d_6^-)$$

$$\begin{cases} 2.5x_1+2.5x_2+5.0x_3+d_1^- -d_1^+ =450 \\ x_1+d_2^- -d_2^+ =30 \\ -x_1+x_2+d_3^- -d_3^+ =30 \\ -x_2+x_3+d_4^- -d_4^+ =0 \\ x_1+d_5^- -d_5^+ =24 \\ x_2+d_6^- -d_6^+ =30 \\ x_1,x_2,x_3,d_i^-,d_i^+ \geqslant 0,\quad (i=1,2,\cdots,6) \end{cases}$$

将求得的部分变量的结果及其含义汇总于表 5-7，在表中未列出的变量，均在满意解时取值为 0。

表 5-7

变　量	含　义	解 1
x_1	晋升到Ⅰ级的人数	24
x_2	晋升到Ⅱ级的人数	30
x_3	新招收Ⅲ级的人数	30
d_1^-	工资总额的结余额	16.5(万元)
d_2^-	Ⅰ级缺编人数	6
d_3^-	Ⅱ级缺编人数	24
d_4^-	Ⅲ级缺编人数	0

例 6　已知有三个产地给四个销地供应某种产品，产销地之间的供需量和单位运价见表 5-8。有关部门在研究调运方案时依次考虑以下七项目标，并规定其相应的优先等级：

P_1——B_4 是重点保证单位，必须全部满足其需要；

P_2——A_3 向 B_1 提供的产量不少于 100；

P_3——每个销地的供应量不小于其需要量的 80%；

P_4——所定调运方案的总运费不超过最小运费调运方案的 10%；

P_5——因路段的问题，尽量避免安排将 A_2 的产品往 B_4；

P_6——给 B_1 和 B_3 的供应率要相同；

P_7——力求总运费最省。

试求满意的调运方案。

表　5-8

产地	销地				产量
	B_1	B_2	B_3	B_4	
A_1	5	2	6	7	300
A_2	3	5	4	6	200
A_3	4	5	2	3	400
销量	200	100	450	250	900/1 000

解　用表上作业法求得最小运费的调运方案见表5-9。这时得最小运费为2 950元，再根据提出的各项目标的要求建立目标规划的模型。模型中 x_{ij} 为由 i 产地调运给 j 销地的产品数。

表　5-9

产地	销地				产量
	B_1	B_2	B_3	B_4	
A_1	200	100			300
A_2	0		200		200
A_3			250	150	400
虚设点				100	100
销量	200	100	450	250	1 000/1 000

供应约束

$$x_{11}+x_{12}+x_{13}+x_{14}\leqslant 300$$
$$x_{21}+x_{22}+x_{23}+x_{24}\leqslant 200$$
$$x_{31}+x_{32}+x_{33}+x_{34}\leqslant 400$$

需求约束

$$x_{11}+x_{21}+x_{31}+d_1^- -d_1^+ =200$$
$$x_{12}+x_{22}+x_{32}+d_2^- -d_2^+ =100$$
$$x_{13}+x_{23}+x_{33}+d_3^- -d_3^+ =450$$
$$x_{14}+x_{24}+x_{34}+d_4^- -d_4^+ =250$$

A_3 向 B_1 提供的产品量不少于100

$$x_{31}+d_5^- -d_5^+ =100$$

每个销地的供应量不小于其需要量的80%

$$x_{11}+x_{21}+x_{31}+d_6^- -d_6^+ =200\times 0.8$$
$$x_{12}+x_{22}+x_{32}+d_7^- -d_7^+ =100\times 0.8$$
$$x_{13}+x_{23}+x_{33}+d_8^- -d_8^+ =450\times 0.8$$
$$x_{14}+x_{24}+x_{34}+d_9^- -d_9^+ =250\times 0.8$$

调运方案的总运费不超过最小运费调运方案的10%

$$\sum_{i=1}^{3}\sum_{j=1}^{4} c_{ij}x_{ij}+d_{10}^{-}-d_{10}^{+}=2\,950(1+10\%)$$

因路段的问题，尽量避免安排将 A_2 的产品运往 B_4

$$x_{24}+d_{11}^{-}-d_{11}^{+}=0$$

给 B_1 和 B_3 的供应率要相同

$$(x_{11}+x_{21}+x_{31})-\frac{200}{450}(x_{13}+x_{23}+x_{33})+d_{12}^{-}-d_{12}^{+}=0$$

力求总运费最省

$$\sum_{i=1}^{3}\sum_{j=1}^{4} c_{ij}x_{ij}+d_{13}^{-}-d_{13}^{+}=2\,950$$

目标函数为

$$\min z=P_1d_4^{-}+P_2d_5^{-}+P_3(d_6^{-}+d_7^{-}+d_8^{-}+d_9^{-})+P_4d_{10}^{+}+P_5d_{11}^{+}+P_6(d_{12}^{-}+d_{12}^{+})+P_7d_{13}^{+}$$

计算结果，得到满意调运方案见表 5-10。

表 5-10

产地	销地				产量
	B_1	B_2	B_3	B_4	
A_1		100		200	300
A_2	90		110		200
A_3	100		250	50	400
虚设点	10		90		100
销 量	200	100	450	250	1 000/1 000

总运费为

$$C=3\times90+4\times100+2\times100+4\times110+2\times250+7\times200+3\times50=3\,360(\text{元})$$

注　　记

用单纯形法求解线性规划问题是令人满意的算法。但在 1972 年 V. Klee 和 G. Minty 构造了一个 n 个变量 $2n$ 个不等式约束的例子。证明用单纯形法求解此线性规划问题时，其计算次数为 2^n。从计算复杂性的标准来看，不是问题规模 n 的多项式函数。凡算法所需计算工作量是问题规模(变量数或方程数)n 的多项式函数时，这算法为有效算法，如高斯消去法的计算量为 $n^3/3$ 的加法，$n^3/2$ 的乘法，$n^2/2$ 的除法。即为 $O(kn^3)$。若某算法的计算量是问题规模的指数式函数 a^n 时，就不是有效算法。因而不能肯定单纯形法是有效算法。于是就产生一问题，能否找到一种解线性规划问题的肯定的多项式算法。1979 年苏联科学院计算中心的哈奇扬(ХАЧИЯН)发表了求解线性规划问题的多项式算法一文，解决了这问题。他提出的椭球算法，并证明了线性规划问题存在多项式算法。但哈奇扬的椭球算法在应用

时，迭代次数比单纯形法还要多，因此理论上意义很大，但实用价值不大。此后卡玛卡尔（N. Karmarkar）又提出了一种新的投影尺度算法参见资料[12]。它本质上是一种内点算法，也属多项式算法。

运筹学的应用与计算机软件的开发紧密联系，规模稍大的运筹学模型是无法用手工计算来求解的。国内目前常用的计算机软件主要有 LINDO，LINGO，WinQSB，MATLAB 和 EXCEL-Risk Solver Platform，Premium Solver Platform 等求解器（solver）。软件商品化进展很快，不时有新版本推出，不需要自己开发。商品品种很多，如 CPLEX，其中用于教学目的软件可从网上免费下载，但有限使用。

对线性规划模型在实践中的模型一般含几百到几千个变量和约束条件，有时变量和约束条件个数可达几十万个到几百万个之多。对这类大规模的模型要用手工建模及输入计算机是不现实的。因此近几年已开发诸如 AMPL（A Mathematical Programming Language），GAMS（The General Algebraic Modeling System）等建模语言，有效地解决了大规模线性规划模型建模和优化计算问题，并还可以用于对这类模型的管理，包括数据存取，将数据转换成模型参数，根据需要修改模型参数，分析模型的解，以及产生决策制定者的内部总结报告和文本等，并与求解器结合使用。介绍以下网址供查阅。

http://www.solver.com/student;

http://www.ampl.com

http://www.theindustrialanalyst.com/analysis_software/linear_programming_tool.htm

http://www-01.ibm.com/software/cn/websphere/ilog/products/ampl/

http://www-01.ibm.com/software/cn/websphere/ilog/products/cplex/index.html

习题

5.1 若用以下表达式作为目标规划的目标函数，试述其逻辑是否正确？

(1) $\max z=d^-+d^+$　　(2) $\max z=d^--d^+$

(3) $\min z=d^-+d^+$　　(4) $\min z=d^--d^+$

5.2 分别用图解法和单纯形法找出以下目标规划问题的满意解。

(1)
$$\min z=P_1(d_1^-+d_1^+)+P_2(2d_2^++d_3^+)$$
$$\begin{cases} x_1-10x_2+d_1^--d_1^+=50 \\ 3x_1+5x_2+d_2^--d_2^+=20 \\ 8x_1+6x_2+d_3^--d_3^+=100 \\ x_1,x_2,d_i^-,d_i^+\geqslant 0,i=1,2,3 \end{cases}$$

(2)
$$\min z=P_1(d_3^++d_4^+)+P_2d_1^++P_3d_2^-+P_4(d_3^-+1.5d_4^-)$$
$$\begin{cases} x_1+x_2+d_1^--d_1^+=40 \\ x_1+x_2+d_2^--d_2^+=100 \\ x_1+d_3^--d_3^+=30 \\ x_2+d_4^--d_4^+=15 \\ x_1,x_2,d_i^-,d_i^+\geqslant 0,i=1,2,3,4 \end{cases}$$

(3)
$$\min z = P_1 d_2^+ + P_1 d_2^- + P_2 d_1^-$$
$$\begin{cases} x_1 + 2x_2 + d_1^- - d_1^+ = 10 \\ 10x_1 + 12x_2 + d_2^- - d_2^+ = 62.4 \\ 2x_1 + x_2 \leqslant 8 \\ x_1, x_2, d_i^-, d_i^+ \geqslant 0;\ i = 1,2 \end{cases}$$

(4)
$$\min z = P_1 d_1^- + P_2 d_2^+ + P_3(5d_3^- + 3d_4^-) + P_4 d_1^+$$
$$\begin{cases} x_1 + x_2 + d_1^- - d_1^+ = 80 \\ x_1 + x_2 + d_2^- - d_2^+ = 90 \\ x_1 + d_3^- - d_3^+ = 70 \\ x_2 + d_4^- - d_4^+ = 45 \\ x_1, x_2, d_i^-, d_i^+ \geqslant 0;\ i = 1,2,3,4 \end{cases}$$

5.3 根据本书第 2 章习题 2.10 给出的某糖果厂生产计划优化的各项数据，若该糖果厂确定生产计划的目标函数为：

P_1——利润不低于某预期值；

P_2——甲，乙，丙三种糖果的原材料比例性满足配方要求；

P_3——充分利用又不超过规定的原材料供应量。

根据上述要求，构建目标规划的数学模型。

5.4 南溪市计划在下一年度预算中购置一批救护车，已知每辆购置价为 20 万元。救护车用于所属 A,B 两个郊区县，各分配 x_A 辆和 x_B 辆。A 县救护站从接到呼叫到出动的响应时间为$(40-3x_A)$分钟；B 县救护站的响应时间为$(50-4x_B)$分钟。该市确定如下优先目标：

P_1——用于救护车的购置费不超过 400 万元；

P_2——A 县的响应时间不超过 8 分钟；

P_3——B 县的响应时间不超过 8 分钟。

要求：

(1) 建立目标规划模型，并求出满意解；

(2) 若对优先级目标函数进行调整，将 P_2 调为 P_1，P_3 调为 P_2，P_1 调为 P_3。试重新构建目标规划的数学模型，并找出新的满意解。

5.5 某商标的酒是用三种等级的酒兑制而成。若这三种等级的酒每天供应量和单位成本见表 5-11。

表 5-11

等 级	日供应量(kg)	成本(元/kg)
Ⅰ	1 500	6
Ⅱ	2 000	4.5
Ⅲ	1 000	3

设该种牌号酒有三种商标（红、黄、蓝），各种商标的酒对原料酒的混合比及售价，见表 5-12。决策者规定：首先必须严格按规定比例兑制各商标的酒；其次是获利最大；再次是红商标的酒每天至少生产 2 000kg，试列出目标规划的数学模型。

表　5-12

商　标	兑制要求/%	售价/(元/kg)
红	Ⅲ少于 10　　Ⅰ多于 50	5.5
黄	Ⅲ少于 70　　Ⅰ多于 20	5.0
蓝	Ⅲ少于 50　　Ⅰ多于 10	4.8

参考资料

[1] Dantzig G B. *Linear Programming 1: Introduction*. Springer, New York, 1997

[2] Dantzig G B. *Linear Programming 2: Theory and Extensions*. Springer, New York, 2003

[3] Vanderbei R J. *Linear Programming: Foundations and Extensions*. 3rd ed., Springer, New York, 2008

[4] 裘宗沪. 解线性规划的单纯形算法中避免循环的几种方法. 数学的实践与认识，1978 年，第 4 期

[5] Schrage L. *Optimization Modeling with LINGO*, LINDO System Press, Chicago, IL, 2008

[6] Bazaraa M S. *Linear Programming and Network Flow*. Second ed. John Wiley & Sons, 1990

[7] Williams H P. *Model Building in Mathematical Programming*. Fourth ed. Wiley, 1999

[8] J. P. Ignizio. *Goal Programming and Extensions*. D. C. Heath and Company, 1976

[9] Hamdy A Taha. *Operations Research—An Introduction*. Eighth ed. Pearson Education Inc. 2007

[10] Cliff Ragsdale. *Spreadsheet Modeling and Decision Analysis: A Practical Introduction to Management Science*. 6th ed., Virginia Polytechnic Institute and State University *College Bookstore* 2011

[11] 越民义. 椭球算法介绍。运筹学杂志，1983 年，第 1 期

[12] N. Karmarkar. *A New Polynomial Time Argorithm for Linear Progamming*. Proceedings of 16th Annual ACM Symposium on Theory of Computing. 1984

[13] Frederick S. Hillier 等著. 运筹学导论（第 9 版）. 北京：清华大学出版社，2010

Part 3

第 3 篇

整数线性规划

CHAPTER 6

第6章

整数线性规划

6.1 整数线性规划问题的提出

在前面讨论的线性规划问题中，有些最优解可能是分数或小数，事实上，线性规划是连续变量的线性优化问题。但在实际中，常有要求解答必须是整数的情形(称为整数解)。例如，所求解是机器的台数、完成工作的人数或装货的车数等，分数或小数的解答就不合要求。为了满足整数解的要求，初看起来，似乎只要把已得到的带有分数或小数的解经过"舍入化整"就可以了。但这常常是不行的，因为化整后不见得是可行解；或虽是可行解，但不一定是最优解。因此，对求最优整数解的问题，有必要另行研究。我们称这样的问题为整数线性规划(integer linear programming)，简称 ILP，整数线性规划是最近几十年来发展起来的数学规划论中的一个分支。

整数线性规划中如果所有的变量都限制为(非负)整数，就称为纯整数线性规划(pure integer linear programming)或称为全整数线性规划(all integer linear programming)；如果仅一部分变量限制为整数，则称为混合整数线性规划(mixed integer linear programming)。整数线性规划的一种特殊情形是 0-1 规划，它的变量取值仅限于 0 或 1。本章最后讲到的指派问题就是一个 0-1 规划问题。

现举例说明用前述单纯形法求得的解不能保证是整数最优解。

例 1 某厂拟用集装箱托运甲乙两种货物，每箱的体积、重量、可获利润以及托运所受限制如表 6-1 所示。问两种货物各托运多少箱，可使获得利润为最大?

表 6-1

货　物	体积/(m^3/箱)	重量/(100kg/箱)	利润/(百元/箱)
甲	5	2	20
乙	4	5	10
托运限制	$24m^3$	1 300kg	

现在我们解这个问题，设 x_1，x_2 分别为甲、乙两种货物的托运箱数(当然都是非负整数)。这是一个(纯)整数规划问题，用数学式可表示为

$$\max z=20x_1+10x_2 \quad ①$$

$$\begin{cases}5x_1+4x_2\leqslant 24 & ②\\ 2x_1+5x_2\leqslant 13 & ③\\ x_1,x_2\geqslant 0 & ④\\ x_1,x_2 \text{ 为整数} & ⑤\end{cases} \quad (6\text{-}1)$$

它和线性规划问题的区别仅在于最后的条件⑤。现在我们暂不考虑这一条件，即解式①～④(以后我们称这样的问题为与原问题相应的线性规划问题)，很容易求得最优解为

$$x_1=4.8,x_2=0,\max z=96$$

但 x_1 是托运甲种货物的箱数，现在它不是整数，所以不合条件⑤的要求。

是不是可以把所得的非整数的最优解经过“化整”就可得到合于条件⑤的整数最优解呢？如将($x_1=4.8,x_2=0$)凑整为($x_1=5,x_2=0$)，这样就破坏了条件②(关于体积的限制)，因而它不是可行解；如将($x_1=4.8,x_2=0$)舍去尾数 0.8，变为($x_1=4,x_2=0$)，这当然满足各约束条件，因而是可行解，但不是最优解，因为

当 $x_1=4,x_2=0$ 时 $z=80$，

但当 $x_1=4,x_2=1$(这也是可行解)时，$z=90$。

本例还可以用图解法来说明。见图 6-1。非整数的最优解在 $C(4.8,0)$点达到。图中画“＋”号的点表示可行的整数解，可见整数线性规划问题的可行域是某相应线性规划可行域中的整数点集(或称格点集)。凑整的(5,0)点不在可行域内，而 C 点又不合于条件⑤。为了满足题中要求，表示目标函数的 z 的等值线必须向原点(即向可行域内部方向)平行移动，直到第一次遇到带“＋”号的 B 点(4,1)为止。这样，z 的等值线就由 $z=96$ 变到 $z=90$，它们的差值

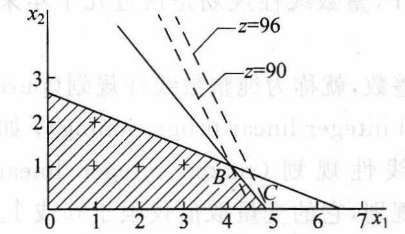


图 6-1

$$\Delta z=96-90=6$$

表示利润的降低，这是由于变量的不可分性(装箱)所引起的。

由上例看出，将其相应的线性规划的最优解“化整”来解原整数规划，虽是最容易想到的，但常常得不到整数规划的最优解，甚至根本不是可行解。因此有必要对整数规划的解法进行专门研究。

6.2 分支定界解法

在求解整数规划时，如果可行域是有界的，首先容易想到的方法就是穷举变量的所有可行的整数组合，就像在图 6-1 中画出所有“＋”号的点那样，然后比较它们的目标函数值以定出最优解。对于小规模的问题，变量数很少，可行的整数组合数也是很小时，这个方法是可行的，也是有效的。在例 1 中，变量只有 x_1 和 x_2；由条件②，x_1 所能取的整数值

为0、1、2、3、4共5个；由条件③，x_2 所能取的整数值为0、1、2共3个，它的组合(不都是可行的)数是 $3\times5=15$(个)，穷举法还是勉强可用的。对于大规模的问题，可行的整数组合数是很大的。例如在本章第5节的指派问题(这也是整数规划)中，将 n 项任务指派 n 个人去完成，不同的指派方案共有 $n!$ 种，当 $n=10$，这个数就超过300万；当 $n=20$，这个数就超过 $20!=2.432\,9\times10^{18}$。如果一一计算，就是用每秒百万次的计算机，也要几万年的工夫，很明显，解这样的题，穷举法是不可取的。

所以我们的方法一般应是仅检查可行的整数组合的一部分，就能定出最优的整数解。**分支定界解法**(branch and bound method)就是其中的一个。分支定界法可用于解纯整数或混合的整数规划问题。在20世纪60年代初由Land Doig和Dakin等人提出。由于该方法灵活且便于用计算机求解，所以现在它已是解整数规划的重要方法。现在大部分整数规划商业软件，如CPLEX和BARON等都是基于分支定界法框架的。

设有最大化的整数规划问题 A，与它相应的线性规划为问题 B，从解问题 B 开始，若其最优解不符合 A 的整数条件，那么 B 的最优目标函数值必是 A 的最优目标函数值 z^* 的上界，记作 $\bar{z}$；而 A 的任意可行解的目标函数值将是 z^* 的一个下界 $\underline{z}$。分支定界法就是将 B 的可行域分成子区域(称为分支)的方法，逐步减小 $\bar{z}$ 和增大 $\underline{z}$，最终求到 z^*。现用下例来说明。

例2　求解 A

$$\max z=40x_1+90x_2 \quad ①$$

$$\begin{cases}9x_1+7x_2\leqslant56 & ②\\ 7x_1+20x_2\leqslant70 & ③\\ x_1,x_2\geqslant0 & ④\\ x_1,x_2\ \text{整数} & ⑤\end{cases} \quad (6\text{-}2)$$

解　先不考虑条件⑤，即解相应的线性规划 B 式①～式④(见图6-2)，得最优解

$$x_1=4.81,\quad x_2=1.82,\quad z_0=356$$

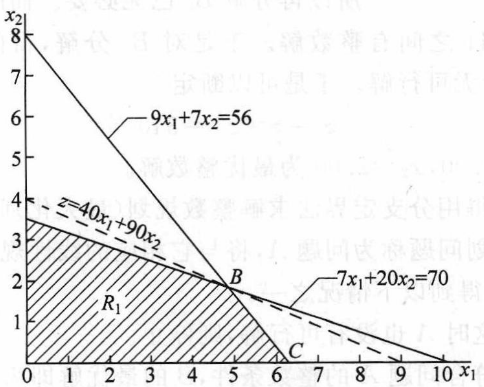


图　6-2

可见它不符合整数条件⑤，但 $z_0=356$ 是问题 A 的最优目标函数值 z^* 的上界，记作 $z_0=\bar{z}$。而 $x_1=0,x_2=0$ 时，显然是问题 A 的一个整数可行解，其相应的 $z=0$，是 z^* 的一个下界，记作 $\underline{z}=0$，即 $0\leqslant z^*\leqslant 356$。

分支定界法的解法，是基于其中非整数变量的解进行分支的，如 x_1，在问题 B 的解中 $x_1=4.81$。基于 x_1，对原问题增加两个约束条件

$$x_1\leqslant 4, x_1\geqslant 5$$

可将原问题分解为两个子问题 B_1 和 B_2（即两支），给每支增加了一个约束条件，如图 6-3 所示。这并不影响问题 A 的可行域，不考虑整数条件解问题 B_1 和 B_2，称此为第一次迭代。得到最优解，如表 6-2 所示。

表 6-2

问题 B_1	问题 B_2
$z_1=349$	$z_2=341$
$x_1=4.00$	$x_1=5.00$
$x_2=2.10$	$x_2=1.57$

显然没有得到全部变量是整数的解。因 $z_1>z_2$，故将 $\bar{z}$ 改为 349，那么必存在最优整数解，得到 z^*，并且

$$0\leqslant z^*\leqslant 349$$

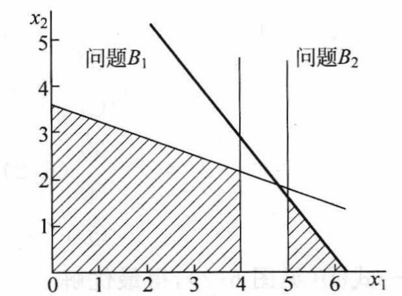


图 6-3

继续对问题 B_1 和 B_2 进行分解，因 $z_1>z_2$，故先分解 B_1 为两支。增加条件 $x_2\leqslant 2$ 者，称为问题 B_3；增加条件 $x_2\geqslant 3$ 者称为问题 B_4。在图 6-3 中再舍去 $x_2>2$ 与 $x_2<3$ 之间的可行域，再进行第二次迭代。解题过程的结果都列在图 6-4 中。可见问题 B_3 的解已都是整数，它的目标函数值 $z_3=340$，可取为 $\underline{z}$，而它大于 $z_4=327$。所以再分解 B_4 已无必要。而问题 B_2 的 $z_2=341$，所以 z^* 可能在 $340\leqslant z^*\leqslant 341$ 之间有整数解。于是对 B_2 分解，得问题 B_5，为非整数解，且 $z_5=308<z_3$，问题 B_6 为无可行解。于是可以断定

$$z_3=\underline{z}=z^*=340$$

问题 B_3 的解 $x_1=4.00,x_2=2.00$ 为最优整数解。

从以上解题过程可得用分支定界法求解整数规划(最大化)问题的步骤为：

将要求解的整数规划问题称为问题 A，将与它相应的线性规划问题称为问题 B。

(1) 解问题 B，可能得到以下情况之一。

① B 没有可行解，这时 A 也没有可行解，则停止。

② B 有最优解，并符合问题 A 的整数条件，B 的最优解即为 A 的最优解，则停止。

③ B 有最优解，但不符合问题 A 的整数条件，记它的目标函数值为 $\bar{z}_0$。

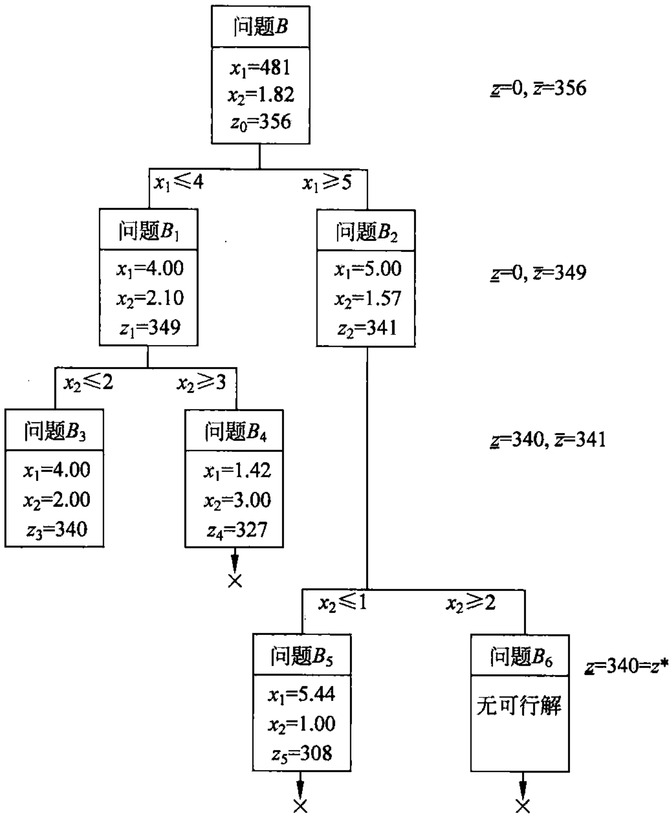


图　6-4

(2) 用观察法找问题 A 的一个整数可行解，如可取 $x_j=0, j=1,\cdots,n$，试探，求得其目标函数值，并记作$\underline{z}$。以 z^* 表示问题 A 的最优目标函数值；这时有

$$\underline{z}\leqslant z^*\leqslant\bar{z}$$

其中，$\underline{z}$ 和$\bar{z}$分别为 z^* 的下界和上界，初始上界即$\bar{z}_0$。

(3) 进行迭代。

第一步：分支，在 B 的最优解中任选一个不符合整数条件的变量 x_j，其值为 b_j，以 $[b_j]$表示小于 b_j 的最大整数。构造两个约束条件

$$x_j\leqslant[b_j] \quad 和 \quad x_j\geqslant[b_j]+1$$

将这两个约束条件，分别加入问题 B，形成两个后继规划问题 B_1 和 B_2。不考虑整数条件求解这两个后继问题。

定界，以每个后继问题为一分支标明求解的结果，与其他问题的解比较，找出最优目标函数值最大者作为新的上界 $\bar{z}$。从已符合整数条件的各分支中，找出目标函数值为最大者作为新的下界$\underline{z}$，若无可行解，$\underline{z}=0$。

第二步：比较与剪支，各分支的最优目标函数值中若有小于$\underline{z}$者，则剪掉这支(用打×表示)，即以后不再考虑了。若大于$\underline{z}$，且不符合整数条件，则重复第一步骤。一直到最后得到$z^*=\underline{z}$为止，得最优整数解x_j^*，$j=1,\cdots,n$。

用分支定界法可解纯整数规划问题和混合整数规划问题。它比穷举法优越。因为它仅在一部分可行解的整数解中寻求最优解，计算量比穷举法小。若变量数目很大，其计算工作量也是相当可观的。

6.3 割平面解法

与分支定界法的相同点是，割平面解法也是将求解整数线性规划的问题化为一系列普通线性规划问题求解。

割平面解法的思路是：首先不考虑变量x_i是整数这一条件，仍然先解其相应的线性规划，若得到非整数的最优解，则增加能割去非整数解的线性约束条件(用几何术语，称为割平面)，使得由原可行域中切割掉一部分，切割掉的部分只包含非整数解，即没有切割掉任何整数可行解。本方法将指出怎样找到适当的割平面(不见得一次就找到)，使切割后最终得到这样的可行域，它的一个有整数坐标的极点恰好是问题的最优解。这个方法是R. E. Gomory提出来的，所以又称为Gomory的割平面法。以下只讨论纯整数规划的情形，现举例说明。

例3 求解

$$\max z=x_1+x_2 \quad ①$$

$$\begin{cases}-x_1+x_2\leqslant 1 & ②\\ 3x_1+x_2\leqslant 4 & ③\\ x_1,x_2\geqslant 0 & ④\\ x_1,x_2\text{整数} & ⑤\end{cases} \quad (6\text{-}3)$$

如不考虑条件⑤，容易求得相应的线性规划的最优解

$$x_1=\frac{3}{4},x_2=\frac{7}{4},\max z=\frac{10}{4}$$

它就是图6-5中域R的极点A，但不合于整数条件。现设想，如能找到像CD那样的直线去切割域R(图6-6)，去掉三角形域ACD，那么具有整数坐标的C点(1,1)就是域R'的一个极点，如在域R'上求解式①～式④，而得到的最优解又恰巧在C点就得到原问题的整数解，所以解法的关键就是怎样构造一个这样的“割平面”CD，尽管它可能不是唯一的，也可能不是一步能求到的。下面仍就本例说明。

在原问题的前两个不等式中增加非负松弛变量x_3、x_4，使两式变成等式约束

$$\begin{cases}-x_1+x_2+x_3\quad\ =1 & ⑥\\ 3x_1+x_2+\quad x_4=4 & ⑦\end{cases}$$

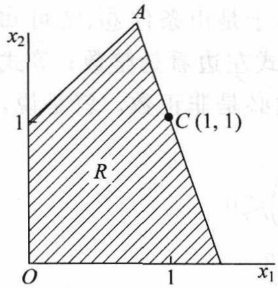


图　6-5

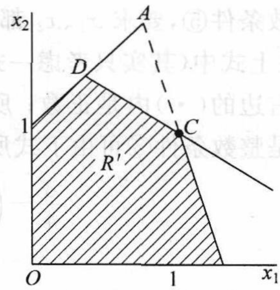


图　6-6

不考虑条件⑤，用单纯形表解题，见表 6-3。

表　6-3

	c_j			1	1	0	0
	C_B	X_B	b	x_1	x_2	x_3	x_4
初始计算表	0	x_3	1	−1	1	1	0
	0	x_4	4	3	1	0	1
	c_j-z_j		0	1	1	0	0
最终计算表	1	x_1	3/4	1	0	−1/4	1/4
	1	x_2	7/4	0	1	3/4	1/4
	c_j-z_j		−5/2	0	0	−1/2	−1/2

从表 6-3 的最终计算表中，得到非整数的最优解

$$x_1=\frac{3}{4},x_2=\frac{7}{4},x_3=x_4=0,\max z=\frac{5}{2}$$

不能满足整数最优解的要求。考虑其中的非整数变量，可以最终计算表中得到相应的关系式

$$x_1-\frac{1}{4}x_3+\frac{1}{4}x_4=\frac{3}{4}$$

$$x_2+\frac{3}{4}x_3+\frac{1}{4}x_4=\frac{7}{4}$$

将系数和常数项都分解成整数和非负真分数两部分之和

$$(1+0)x_1+\left(-1+\frac{3}{4}\right)x_3+\frac{1}{4}x_4=0+\frac{3}{4}$$

$$x_2+\frac{3}{4}x_3+\frac{1}{4}x_4=1+\frac{3}{4}$$

然后将整数部分与分数部分分开，移到等式左右两边，得到

$$x_1-x_3=\frac{3}{4}-\left(\frac{3}{4}x_3+\frac{1}{4}x_4\right)$$

$$x_2-1=\frac{3}{4}-\left(\frac{3}{4}x_3+\frac{1}{4}x_4\right)$$

现考虑整数条件⑤，要求 x_1、x_2 都是非负整数，于是由条件⑥、⑦可知 x_3、x_4 也都是非负整数①。在上式中(其实只考虑一式即可)从等式左边看是整数；等式右边也应是整数。但在等式右边的(·)内是正数；所以等式右边必是非正数。就是说，右边的整数值最大是零。于是整数条件⑤可由下式所代替；

$$\frac{3}{4}-\left(\frac{3}{4}x_3+\frac{1}{4}x_4\right)\leqslant 0$$

即

$$-3x_3-x_4\leqslant -3 \qquad ⑧$$

这就得到一个切割方程(或称为切割约束)，将它作为增加的约束条件，再解例 3。

引入松弛变量 x_5，得到等式

$$-3x_3-x_4+x_5=-3$$

将这个新的约束方程加到表 6-3 的最终计算表，得表 6-4。

从表 6-4 的 $\boldsymbol{b}$ 列中可看到，这时得到的是非可行解，于是需要用对偶单纯形法继续进行计算。选择 x_5 为换出变量，计算

$$\theta=\min_j\left(\frac{c_j-z_j}{\alpha_{lj}}\middle|\alpha_{lj}<0\right)=\min\left(\frac{-\frac{1}{2}}{-3},\frac{-\frac{1}{2}}{-1}\right)=\frac{1}{6}$$

表 6-4

c_j			1	1	0	0	0
$\boldsymbol{C}_B$	$\boldsymbol{X}_B$	$\boldsymbol{b}$	x_1	x_2	x_3	x_4	x_5
1	x_1	3/4	1	0	−1/4	1/4	0
1	x_2	7/4	0	1	3/4	1/4	0
0	x_5	−3	0	0	−3	−1	1
c_j-z_j		−5/2	0	0	−1/2	−1/2	0

将 x_3 作为换入变量，再按原单纯形法进行迭代，得表 6-5。

表 6-5

c_j			1	1	0	0	0
$\boldsymbol{C}_B$	$\boldsymbol{X}_B$	$\boldsymbol{b}$	x_1	x_2	x_3	x_4	x_5
1	x_1	1	1	0	0	1/3	−1/12
1	x_2	1	0	1	0	0	1/4
0	x_3	1	0	0	1	1/3	−1/3
c_j-z_j		−2	0	0	0	−1/3	−1/6

① 这一点对以下推导是必要的，如不都是整数，则应在引入 x_3、x_4 之前乘以适当常数，使之都是整数。

由于 x_1、x_2 的值已都是整数，解题已完成。

注意：新得到的约束条件⑧

$$-3x_3-x_4\leqslant -3$$

如用 x_1、x_2 表示，由式⑥、式⑦得

$$3(1+x_1-x_2)+(4-3x_1-x_2)\geqslant 3$$

$$x_2\leqslant 1$$

这就是(x_1,x_2)平面内形成新的可行域，即包括平行于 x_1 轴的直线 $x_2=1$ 和这直线下的可行区域，整数点也在其中，没有被切割掉。直观地表示在图 6-7 中。但从解题过程来看，这一步是不必要的。

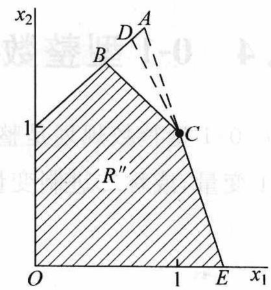


图 6-7

现把求一个切割方程的步骤归纳为：

(1) 令 x_i 是相应线性规划最优解中为分数值的一个基变量，由单纯形表的最终表得到

$$x_i+\sum_k \alpha_{ik}x_k=b_i \tag{6-4}$$

其中，$i\in Q$（Q 指构成基变量号码的集合）；

$k\in K$（K 指构成非基变量号码的集合）。

(2) 将 b_i 和 α_{ik} 都分解成整数部分 N 与非负真分数 f 之和，即

$$b_i=N_i+f_i,\quad \text{其中 } 0<f_i<1$$

$$\alpha_{ik}=N_{ik}+f_{ik},\quad \text{其中 } 0\leqslant f_{ik}<1 \tag{6-5}$$

而 N 表示不超过 b 的最大整数。例如，

若 $b=2.35$，　　则 $N=2, f=0.35$

若 $b=-0.45$，　　则 $N=-1, f=0.55$

代入式(6-4)得

$$x_i+\sum_k N_{ik}x_k-N_i=f_i-\sum_k f_{ik}x_k \tag{6-6}$$

(3) 现在提出变量（包括松弛变量，参阅例 3 的注）为整数的条件（当然还有非负的条件），这时，上式由左边看必须是整数，但由右边看，因为 $0<f_i<1$，所以不能为正，即

$$f_i-\sum_k f_{ik}x_k\leqslant 0 \tag{6-7}$$

这就是一个切割方程。

由式(6-4)、式(6-6)和式(6-7)可知：

① 切割方程式(6-7)真正进行了切割，至少把非整数最优解这一点割掉了。

② 没有割掉整数解，这是因为相应的线性规划的任意整数可行解都满足式(6-7)的缘故。

Gomory 的切割法自 1958 年被提出后，即引起人们广泛的注意，至今还在不断地改进。

6.4 0-1 型整数线性规划

0-1 型整数规划是整数规划中的特殊情形,它的变量 x_i 仅取值 0 或 1。这时 x_i 称为 0-1 变量,或称二进制变量①。x_i 仅取值 0 或 1 这个条件可由下述约束条件所代替。

$$x_i \leqslant 1$$
$$x_i \geqslant 0, \text{整数}$$

它和一般整数规划的约束条件形式是一致的。如果变量 x_i 不是仅取值 0 或 1,而是可取其他范围的非负整数,这时可利用二进制的记数法将它用若干个 0-1 变量来代替。例如,在给定的问题中,变量 x 可取 0 与 10 之间的任意整数时,令

$$x = 2^0 x_0 + 2^1 x_1 + 2^2 x_2 + 2^3 x_3$$

则 x 就可用 4 个 0-1 变量 x_0, x_1, x_2, x_3 来代替,因此 0-1 变量也称二进制变量。

在实际问题中,如果引入 0-1 变量,就可以把有各种情况需要分别讨论的线性规划问题统一在一个问题中讨论了。在本节我们先介绍引入 0-1 变量的实际问题,再研究解法。

6.4.1 引入 0-1 变量的实际问题

1. 投资场所的选定——相互排斥的计划

例 4 某公司拟在市东、西、南三区建立门市部。拟议中有 7 个位置(点)A_i $(i=1, 2, \cdots, 7)$可供选择。规定:

在东区,由 A_1, A_2, A_3 三个点中至多选两个;

在西区,由 A_4, A_5 两个点中至少选一个;

在南区,由 A_6, A_7 两个点中至少选一个。

如选用 A_i 点,设备投资估计为 b_i 元,每年可获利润估计为 c_i 元,但投资总额不能超过 B 元。问应选择哪几个点可使年利润为最大?

解题时先引入 0-1 变量 $x_i (i=1,2,\cdots,7)$

令

$$x_i = \begin{cases} 1, \text{当 } A_i \text{ 点被选用}, \\ 0, \text{当 } A_i \text{ 点没被选用}。 \end{cases} \quad i=1,2,\cdots,7$$

于是问题可表示为

---

① 如果变量 x_i 不是仅取值 0 或 1,而是可取其他范围的非负整数,这时可利用二进制的记数法将它用若干个 0-1 变量来代替。例如,在给定的问题中,变量 x 可任取 0 与 10 之间的任意整数时,令

$$x = 2^0 x_0 + 2^1 x_1 + 2^2 x_2 + 2^3 x_3$$

这时,x 就可用 4 个 0-1 变量 x_0, x_1, x_2, x_3 来代替。

$$\max z=\sum_{i=1}^{7} c_i x_i$$

$$\begin{cases}\sum_{i=1}^{7} b_i x_i \leqslant B \\ x_1+x_2+x_3 \leqslant 2 \\ x_4+x_5 \geqslant 1 \\ x_6+x_7 \geqslant 1 \\ x_i=0 \text{ 或 } 1\end{cases} \tag{6-8}$$

2. 相互排斥的约束条件

在本章开始的例 1 中，关于运货的体积限制为

$$5x_1+4x_2 \leqslant 24 \tag{6-9}$$

今设运货有车运和船运两种方式，上面的条件系用车运时的限制条件，如用船运时关于体积的限制条件为

$$7x_1+3x_2 \leqslant 45 \tag{6-10}$$

这两条件是互相排斥的。为了统一在一个问题中，引入 0-1 变量 y，令

$$y=\begin{cases}0, & \text{当采取车运方式} \\ 1, & \text{当采取船运方式}\end{cases}$$

于是式(6-9)和式(6-10)可由下述的条件式(6-11)和式(6-12)来代替

$$5x_1+4x_2 \leqslant 24+yM \tag{6-11}$$

$$7x_1+3x_2 \leqslant 45+(1-y)M \tag{6-12}$$

其中，M 是充分大的数。读者可以验证，当 $y=0$ 时，式(6-11)就是式(6-9)，而式(6-12)自然成立，因而是多余的。当 $y=1$ 时式(6-12)就是式(6-10)，而式(6-11)是多余的。引入的变量 y 不必出现在目标函数内，即认为在目标函数式内 y 的系数为 0。

如果有 m 个互相排斥的约束条件（≤型）

$$a_{i1}x_1+a_{i2}x_2+\cdots+a_{in}x_n \leqslant b_i, \quad i=1,2,\cdots,m$$

为了保证这 m 个约束条件只有一个起作用，可引入 m 个 0-1 变量 $y_i(i=1,2,\cdots,m)$ 和一个充分大的常数 M，而下面这一组 $m+1$ 个约束条件

$$a_{i1}x_1+a_{i2}x_2+\cdots+a_{in}x_n \leqslant b_i+y_iM, \quad i=1,2,\cdots,m \tag{6-13}$$

$$y_1+y_2+\cdots+y_m=m-1 \tag{6-14}$$

就合于上述的要求。这是因为，由于式(6-14)，m 个 y_i 中只有一个能取 0 值，设$y_{i^*}=0$，代入式(6-13)，就只有 $i=i^*$ 的约束条件起作用，而别的式子都是多余的。

3. 关于固定费用的问题(fixed cost problem)

在讨论线性规划时，有些问题是要求使成本为最小。那时总设固定成本为常数，并在线性规划的模型中不必明显列出。而有些固定费用（固定成本）的问题不能用一般线性规划来描述，但可改变为混合整数规划来解决，见下例。

例 5 某工厂为了生产某种产品，有几种不同的生产方式可供选择，如选定投资高的

生产方式(选购自动化程度高的设备),由于产量大,因而分配到每件产品的变动成本就降低;反之,如选定投资低的生产方式,将来分配到每件产品的变动成本可能增加,所以必须全面考虑。今设有三种方式可供选择,令

x_j 表示采用第 j 种方式时的产量;

c_j 表示采用第 j 种方式时每件产品的变动成本;

k_j 表示采用第 j 种方式时的固定成本。

为了说明成本的特点,暂不考虑其他约束条件。采用各种生产方式的总成本分别为

$$P_j=\begin{cases}k_j+c_jx_j, & \text{当 } x_j>0 \\ 0, & \text{当 } x_j=0\end{cases} \quad j=1,2,3$$

在构成目标函数时,为了统一在一个问题中讨论,现引入 0-1 变量 y_j,令

$$y_j=\begin{cases}1, & \text{当采用第 } j \text{ 种生产方式,即 } x_j>0 \text{ 时,} \\ 0, & \text{当不采用第 } j \text{ 种生产方式,即 } x_j=0 \text{ 时。}\end{cases} \tag{6-15}$$

于是目标函数

$$\min z=(k_1y_1+c_1x_1)+(k_2y_2+c_2x_2)+(k_3y_3+c_3x_3)$$

式(6-15)这个规定可由下述 3 个线性约束条件表示:

$$x_j \leqslant y_jM, \quad j=1,2,3 \tag{6-16}$$

其中,M 是个充分大的常数。式(6-16)说明,当 $x_j>0$ 时 y_j 必须为 1;当 $x_j=0$ 时只有 y_j 为 0 时才有意义,所以式(6-16)完全可以代替式(6-15)。

6.4.2 0-1 型整数规划的解法

解 0-1 型整数规划最容易想到的方法,和一般整数规划的情形一样,就是穷举法,即检查变量取值为 0 或 1 的每一种组合,比较目标函数值以求得最优解,这就需要检查变量取值的 2^n 个组合。对于变量个数 n 较大(例如 $n>10$),这几乎是不可能的。因此常设计一些方法,只检查变量取值的组合的一部分,就能求到问题的最优解。这样的方法称为隐枚举法(implicit enumeration),分支定界法也是一种隐枚举法。另外,还有拉格朗日松弛法。下面举例介绍隐枚举法。

下面举例说明一种解 0-1 型整数规划的隐枚举法。

例 6

$$\max z=3x_1-2x_2+5x_3$$

$$\begin{cases}x_1+2x_2-x_3\leqslant 2 & ① \\ x_1+4x_2+x_3\leqslant 4 & ② \\ x_1+x_2\leqslant 3 & ③ \\ 4x_1+x_3\leqslant 6 & ④ \\ x_1,x_2,x_3=0 \text{ 或 } 1 & ⑤\end{cases} \tag{6-17}$$

解题时先通过试探的方法找一个可行解,容易看出$(x_1,x_2,x_3)=(1,0,0)$就是合于条件①~④的,算出相应的目标函数值 $z=3$。

对于极大化问题，当然希望 $z \geqslant 3$，于是在式①之前增加一个约束条件

$$3x_1-2x_2+5x_3\geqslant 3 \qquad ◎$$

增加的这个约束条件称为过滤的条件(filtering constraint)。这样，原问题的线性约束条件就变成 5 个。用全部枚举的方法，3 个变量共有 $2^3=8$ 个解，原来 4 个约束条件，共需 32 次运算。现在增加了过滤条件◎，如按下述方法进行，就可减少运算次数。将 5 个约束条件按见◎～④顺序排好(见表 6-6)，对每个解，依次代入约束条件左侧，求出数值，看是否适合不等式条件，如某一条件不适合，同行以下各条件就不必再检查，因而就减少了运算次数。本例计算过程如表 6-6，实际只作 24 次运算。

于是求得最优解　　$(x_1,x_2,x_3)=(1,0,1)$　　$\max z=8$

在计算过程中，若遇到 z 值已超过条件◎右边的值，应改变条件◎，使右边为迄今为止最大者，然后继续作。例如，当检查点(0,0,1)时因 $z=5(>3)$，所以应将条件◎换成

$$3x_1-2x_2+5x_3\geqslant 5 \qquad ◎$$

这种对过滤条件的改进，更可以减少计算量。

表 6-6

点	条件					满足条件？是(√)否(×)	z 值
	◎	①	②	③	④		
(0,0,0)	0					×	
(0,0,1)	5	−1	1	0	1	√	5
(0,1,0)	−2					×	
(0,1,1)	3	1	5			×	
(1,0,0)	3	1	1	1	0	√	3
(1,0,1)	8	0	2	1	1	√	8
(1,1,0)	1					×	
(1,1,1)	6	2	6			×	

注意：一般常重新排列 x_i 的顺序使目标函数中 x_i 的系数是递增(不减)的，如在上例中，改写 $z=3x_1-2x_2+5x_3=-2x_2+3x_1+5x_3$。

因为−2,3,5 是递增的，变量(x_2,x_1,x_3)也按下述顺序取值：(0,0,0),(0,0,1),(0,1,0),(0,1,1),…，这样，最优解容易比较早的发现。再结合过滤条件的改进，更可使计算简化。在上例中

$$\max z=-2x_2+3x_1+5x_3$$

$$\begin{cases}-2x_2+3x_1+5x_3\geqslant 3 & ◎\\ 2x_2+x_1-x_3\leqslant 2 & ①\\ 4x_2+x_1+x_3\leqslant 4 & ②\ (6\text{-}18)\\ x_2+x_1\leqslant 3 & ③\\ 4x_1+x_3\leqslant 6 & ④\end{cases}$$

解题时按下述步骤进行(见表 6-7)

表 6-7(a)

点 (x_2,x_1,x_3)	条件					是否满足条件	z 值
	◎	①	②	③	④		
(0,0,0)	0					×	
(0,0,1)	5	−1	1	0	1	√	5

表 6-7(b)

点 (x_2,x_1,x_3)	条件					是否满足条件	z 值
	◎′	①	②	③	④		
(0,1,0)	3					×	
(0,1,1)	8	0	2	1	1	√	8

改进过滤条件,用

$$-2x_2+3x_1+5x_3\geqslant 5 \qquad ◎'$$

代替式◎,继续进行。
再改进过滤条件,用

$$2x_2+3x_1+5x_3\geqslant 8 \qquad ◎''$$

代替式◎′,再继续进行。至此,z 值已不能改进,即得到最优解,解答如前,但计算已简化。

表 6-7(c)

点 (x_2,x_1,x_3)	条件					是否满足条件	z 值
	◎″	①	②	③	④		
(1,0,0)	2					×	
(1,0,1)	3					×	
(1,1,0)	1					×	
(1,1,1)	6					×	

6.5 指派问题

在生活中经常遇到这样的问题,某单位需完成 n 项任务,恰好有 n 个人可承担这些任务。由于每人的专长不同,各人完成任务不同(或所费时间),效率也不同。于是产生应指派哪个人去完成哪项任务,使完成 n 项任务的总效率最高(或所需总时间最小)的问题。这类问题称为指派问题或分派问题(assignment problem)。

例 7 有一份中文说明书,需译成英、日、德、俄四种文字。分别记作 E、J、G、R。现有甲、乙、丙、丁四人。他们将中文说明书翻译成不同语种的说明书所需时间如表 6-8 所

示。问应指派何人去完成何工作，使所需总时间最少？

表 6-8

人员	任务			
	E	J	G	R
甲	2	15	13	4
乙	10	4	14	15
丙	9	14	16	13
丁	7	8	11	9

类似有：有 n 项加工任务，怎样指派到 n 台机床上分别完成的问题；有 n 条航线，怎样指定 n 艘船去航行的问题……对应每个指派问题，需有类似表 6-8 那样的数表，称为效率矩阵或系数矩阵，其元素 $c_{ij}>0(i,j=1,2,\cdots,n)$ 表示指派第 i 人去完成第 j 项任务时的效率（或时间、成本等）。解题时需引入变量 x_{ij}；其取值只能是 1 或 0。并令

$$x_{ij}=\begin{cases}1 & \text{当指派第 } i \text{ 人去完成第 } j \text{ 项任务}\\ 0 & \text{当不指派第 } i \text{ 人去完成第 } j \text{ 项任务}\end{cases}$$

当问题要求极小化时数学模型是

$$\min z=\sum_i\sum_j c_{ij}x_{ij} \qquad ①$$

$$\begin{cases}\sum\limits_i x_{ij}=1, & j=1,2,\cdots,n \qquad ②\\ \sum\limits_j x_{ij}=1, & i=1,2,\cdots,n \qquad ③\\ x_{ij}=1 \text{ 或 } 0 & \qquad ④\end{cases} \qquad (6\text{-}19)$$

约束条件②说明第 j 项任务只能由 1 人去完成；约束条件③说明第 i 人只能完成 1 项任务。满足约束条件②～④的可行解 x_{ij} 也可写成表格或矩阵形式，称为解矩阵。如例 7 的一个可行解矩阵是

$$(x_{ij})=\begin{bmatrix}0&1&0&0\\0&0&1&0\\1&0&0&0\\0&0&0&1\end{bmatrix}$$

解矩阵 (x_{ij}) 中各行各列的元素之和都是 1，但这不是最优解。

指派问题是 0-1 规划的特例，也是运输问题的特例；即 $n=m, a_j=b_i=1$。当然可用整数规划，0-1 规划或运输问题的解法去求解，但这就如同用单纯形法求解运输问题一样是不合算的。利用指派问题的特点可有更简便的解法。

指派问题的最优解有这样性质，若从系数矩阵 (c_{ij}) 的一行（列）各元素中分别减去该

行(列)的最小元素,得到新矩阵(b_{ij}),那么以(b_{ij})为系数矩阵求得的最优解和用原系数矩阵求得的最优解相同。

利用这个性质,可使原系数矩阵变换为含有很多0元素的新系数矩阵,而最优解保持不变,在系数矩阵(b_{ij})中,我们关心位于不同行不同列的0元素,以下简称为独立的0元素。若能在系数矩阵(b_{ij})中找出n个独立的0元素;则令解矩阵(x_{ij})中对应这n个独立的0元素的元素取值为1,其他元素取值为0。将其代入目标函数中得到$z_b=0$,它一定是最小。这就是以(b_{ij})为系数矩阵的指派问题的最优解。也就得到了原问题的最优解。

库恩(W. W. Kuhn)于1955年提出了指派问题的解法,他引用了匈牙利数学家康尼格(D. König)一个关于矩阵中0元素的定理:系数矩阵中独立0元素的最多个数等于能覆盖所有0元素的最少直线数。这解法称为匈牙利法。以后在方法上虽有不断改进,但仍沿用这名称。以下用例7来说明指派问题的匈牙利解法。

第一步:使指派问题的系数矩阵经变换,在各行各列中都出现0元素。

(1) 从系数矩阵的每行元素减去该行的最小元素;

(2) 再从所得系数矩阵的每列元素中减去该列的最小元素。

若某行(列)已有0元素,那就不必再减了。例7的计算为

$$
(c_{ij})=\begin{bmatrix} 2 & 15 & 13 & 4 \\ 10 & 4 & 14 & 15 \\ 9 & 14 & 16 & 13 \\ 7 & 8 & 11 & 9 \end{bmatrix}\begin{matrix} \min \\ 2 \\ 4 \\ 9 \\ 7 \end{matrix} \longrightarrow \begin{bmatrix} 0 & 13 & 11 & 2 \\ 6 & 0 & 10 & 11 \\ 0 & 5 & 7 & 4 \\ 0 & 1 & 4 & 2 \end{bmatrix} \longrightarrow \begin{bmatrix} 0 & 13 & 7 & 0 \\ 6 & 0 & 6 & 9 \\ 0 & 5 & 3 & 2 \\ 0 & 1 & 0 & 0 \end{bmatrix}=(b_{ij})
$$

$$
\qquad\qquad\qquad\qquad 4 \quad 2 \quad \min
$$

第二步:进行试指派,以寻求最优解。为此,按以下步骤进行。

经第一步变换后,系数矩阵中每行每列都已有了0元素;但需找出n个独立的0元素。若能找出,就以这些独立0元素对应解矩阵(x_{ij})中的元素为1,其余为0,这就得到最优解。当n较小时,可用观察法、试探法去找出n个独立0元素。若n较大时,就必须按一定的步骤去找,常用的步骤为:

(1) 从只有一个0元素的行开始,给这个0元素加圈,记作◎。这表示对这行所代表的人,只有一种任务可指派。然后划去◎所在列的其他0元素,记作Φ。这表示这列所代表的任务已指派完,不必再考虑别人了。

(2) 给只有一个0元素列的0元素加圈,记作◎;然后划去◎所在行的其他0元素,记作Φ。

(3) 反复进行(1),(2)两步,直到所有0元素都被圈出和划掉为止。

(4) 若仍有没有划圈的0元素,且同行的0元素至少有两个(表示对这个人可以从两项任务中指派其一)。这可用不同的方案去试探。从剩有0元素最少的行(列)开始,比较这行各0元素所在列中0元素的数目,选择0元素少的那列的这个0元素加圈(表示选择

性多的要“礼让”选择性少的）。然后划掉同行同列的其他 0 元素。可反复进行，直到所有 0 元素都已圈出和划掉为止。

(5) 若◎元素的数目 m 等于矩阵的阶数 n，那么这指派问题的最优解已得到。若 $m<n$，则转入下一步。

现用例 7 的(b_{ij})矩阵，按上述步骤进行运算。按步骤(1)，先给 b_{22} 加圈，然后给 b_{31} 加圈，划掉 b_{11}，b_{41}；按步骤(2)，给 b_{43} 加圈，划掉 b_{44}，最后给 b_{14} 加圈，得到

$$\begin{bmatrix} \Phi & 13 & 7 & ◎ \\ 6 & ◎ & 6 & 9 \\ ◎ & 5 & 3 & 2 \\ \Phi & 1 & ◎ & \Phi \end{bmatrix}$$

可见 $m=n=4$，所以得最优解为

$$(x_{ij})=\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

这表示：指定甲译出俄文，乙译出日文，丙译出英文，丁译出德文。所需总时间最少

$$\min z_b=\sum_i\sum_j b_{ij}x_{ij}=0$$

$$\min z=\sum_i\sum_j c_{ij}x_{ij}=c_{31}+c_{22}+c_{43}+c_{14}=28(\text{小时})$$

例 8 求表 6-9 所示效率矩阵的指派问题的最小解。

表 6-9

人员	任务				
	A	*B*	*C*	*D*	*E*
甲	12	7	9	7	9
乙	8	9	6	6	6
丙	7	17	12	14	9
丁	15	14	6	6	10
戊	4	10	7	10	9

解题时按上述第一步，将这系数矩阵进行变换。

$$\begin{bmatrix} 12 & 7 & 9 & 7 & 9 \\ 8 & 9 & 6 & 6 & 6 \\ 7 & 17 & 12 & 14 & 9 \\ 15 & 14 & 6 & 6 & 10 \\ 4 & 10 & 7 & 10 & 9 \end{bmatrix} \begin{matrix} \min \\ 7 \\ 6 \\ 7 \\ 6 \\ 4 \end{matrix} \longrightarrow \begin{bmatrix} 5 & 0 & 2 & 0 & 2 \\ 2 & 3 & 0 & 0 & 0 \\ 0 & 10 & 5 & 7 & 2 \\ 9 & 8 & 0 & 0 & 4 \\ 0 & 6 & 3 & 6 & 5 \end{bmatrix}$$

经一次运算即得每行每列都有 0 元素的系数矩阵，再按上述步骤运算，得到

$$\begin{bmatrix} 5 & ◎ & 2 & \Phi & 2 \\ 2 & 3 & \Phi & ◎ & \Phi \\ ◎ & 10 & 5 & 7 & 2 \\ 9 & 8 & ◎ & \Phi & 4 \\ \Phi & 6 & 3 & 6 & 5 \end{bmatrix} \qquad ①$$

这里◎的个数 $m=4$，而 $n=5$；所以解题没有完成，这时应按以下步骤继续进行。

第三步：作最少的直线覆盖所有 0 元素，以确定该系数矩阵中能找到最多的独立元素数。为此按以下步骤进行：

(1) 对没有◎的行打√号；

(2) 对已打√号的行中所有含 Φ 元素的列打√号；

(3) 再对打有√号的列中含◎元素的行打√号；

(4) 重复(2)，(3)直到得不出新的打√号的行、列为止；

(5) 对没有打√号的行画一横线，有打√号的列画一纵线，这就得到覆盖所有 0 元素的最少直线数。

令这直线数为 l。若 $l<n$，说明必须再变换当前的系数矩阵，才能找到 n 个独立的 0 元素，为此转第四步：若 $l=n$，而 $m<n$，应回到第二步(4)，另行试探。

在例 8 中，对矩阵①按以下次序进行：

先在第五行旁打√，接着可判断应在第 1 列下打√，接着在第 3 行旁打√。经检查不再能打√了。对没有打√行，画一直线以覆盖 0 元素，已打√的列画一直线以覆盖 0 元素。得

$$\begin{bmatrix} 5 & ◎ & 2 & \Phi & 2 \\ 2 & 3 & \Phi & ◎ & \Phi \\ ◎ & 10 & 5 & 7 & 2 \\ 9 & 8 & ◎ & \Phi & 4 \\ \Phi & 6 & 3 & 6 & 5 \end{bmatrix}\begin{array}{l} \\ \\ \surd \\ \\ \surd \end{array} \qquad ②$$

$\surd$（第 1 列下）

由此可见 $l=4<n$。所以应继续对矩阵②进行变换。转第四步。

第四步：对矩阵②进行变换的目的是增加 0 元素。为此在没有被直线覆盖的部分中找出最小元素。然后在打√行各元素中都减去这最小元素，而在打√列的各元素都加上这最小元素，以保证原来 0 元素不变。这样得到新系数矩阵（它的最优解和原问题相同）。若得到 n 个独立的 0 元素，则已得最优解，否则回到第三步重复进行。

在例 8 的矩阵②中，在没有被覆盖部分（第 3、5 行）中找出最小元素为 2，然后在第 3、5 行各元素分别减去 2，给第 1 列各元素加 2，得到新矩阵③。按第二步，找出所有独立的

0 元素，得到矩阵④。

$$\begin{bmatrix} 7 & 0 & 2 & 0 & 2 \\ 4 & 3 & 0 & 0 & 0 \\ 0 & 8 & 3 & 5 & 0 \\ 11 & 8 & 0 & 0 & 4 \\ 0 & 4 & 1 & 4 & 3 \end{bmatrix} \quad ③$$

$$\begin{bmatrix} 7 & ◎ & 2 & \Phi & 2 \\ 4 & 3 & \Phi & ◎ & \Phi \\ \Phi & 8 & 3 & 5 & ◎ \\ 11 & 8 & ◎ & \Phi & 4 \\ ◎ & 4 & 1 & 4 & 3 \end{bmatrix} \quad ④$$

它具有 n 个独立 0 元素。这就得到了最优解，相应的解矩阵为

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

由解矩阵得最优指派方案

甲—B，乙—D，丙—E，丁—C，戊—A

本例还可以得到另一最优指派方案

甲—B，乙—C，丙—E，丁—D，戊—A

所需总时间为 $\min z=32$

当指派问题的系数矩阵，经过变换得到了同行和同列中都有两个或两个以上 0 元素时。这时可以任选一行(列)中某一个 0 元素，再划去同行(列)的其他 0 元素。这时会出现多重解。

以上讨论限于极小化的指派问题。对极大化的问题，即求

$$\max z=\sum_i\sum_j c_{ij}x_{ij} \tag{6-20}$$

可令

$$b_{ij}=M-c_{ij}$$

其中，M 是足够大的常数(如选 c_{ij} 中最大元素为 M 即可)，这时系数矩阵可变换为

$$\boldsymbol{B}=(b_{ij})$$

这时 $b_{ij}\geqslant 0$，符合匈牙利法的条件。目标函数经变换后，即解

$$\min z'=\sum_i\sum_j b_{ij}x_{ij} \tag{6-21}$$

所得最小解就是原问题的最大解，因为

$$\begin{aligned}\sum_i\sum_j b_{ij}x_{ij} &= \sum_i\sum_j (M-c_{ij})x_{ij}\\ &= \sum_i\sum_j Mx_{ij} - \sum_i\sum_j c_{ij}x_{ij}\\ &= nM - \sum_i\sum_j c_{ij}x_{ij}\end{aligned}$$

因 nM 为常数，所以当 $\sum_i\sum_j b_{ij}x_{ij}$ 取最小时，$\sum_i\sum_j c_{ij}x_{ij}$ 便为最大。

习题

6.1 对下列整数规划问题，问用先解相应的线性规划然后凑整的办法能否求到最优整数解？

(1) $\max z=3x_1+2x_2$

$$\begin{cases}2x_1+3x_2\leqslant 14.5\\ 4x_1+x_2\leqslant 16.5\\ x_1,x_2\geqslant 0\\ x_1,x_2 \text{ 整数}\end{cases}$$

(2) $\max z=3x_1+2x_2$

$$\begin{cases}2x_1+3x_2\leqslant 14\\ 2x_1+x_2\leqslant 9\\ x_1,x_2\geqslant 0\\ x_1,x_2 \text{ 整数}\end{cases}$$

6.2 用分支定界法解

$$\max z=x_1+x_2$$

$$\begin{cases}x_1+\dfrac{9}{14}x_2\leqslant\dfrac{51}{14}\\ -2x_1+x_2\leqslant\dfrac{1}{3}\\ x_1,x_2,\geqslant 0\\ x_1,x_2 \text{ 整数}\end{cases}$$

6.3 用 Gomory 切割法解

(1) $\max z=x_1+x_2$

$$\begin{cases}2x_1+x_2\leqslant 6\\ 4x_1+5x_2\leqslant 20\\ x_1,x_2\geqslant 0\\ x_1,x_2 \text{ 整数}\end{cases}$$

(2) $\max z=3x_1-x_2$

$$\begin{cases}3x_1-2x_2\leqslant 3\\ -5x_1-4x_2\leqslant -10\\ 2x_1+x_2\leqslant 5\\ x_1,x_2\geqslant 0\\ x_1,x_2 \text{ 整数}\end{cases}$$

6.4 某城市的消防总部将全市划分为 11 个防火区，设有 4 个消防(救火)站。图 6-8 表示各防火区域与消防站的位置，其中①②③④表示消防站，1、2、…、11 表示防火区域。根据历史的资料证实，各消防站可在事先规定的允许时间内对所负责的地区的火灾予以消灭。图中虚线即表示各地区由哪个消防站负责(没有虚线连接，就表示不负责)。现在

总部提出：可否减少消防站的数目，仍能同样负责各地区的防火任务？如果可以，应当关闭哪个？

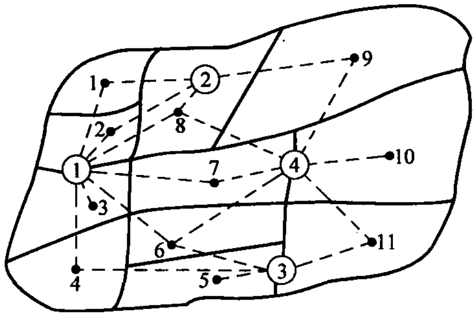


图 6-8

提示：对每个消防站定义一个 0-1 变量 x_i，

令

$$x_j=\begin{cases}1, & \text{当某防火区域可由第 } j \text{ 消防站负责时,}\\ 0, & \text{当某防火区域不由第 } j \text{ 消防站负责时,}\end{cases}$$

$$j=1,2,3,4$$

然后对每个防火区域列一个约束条件。

6.5 某大型企业每年需要进行多种类型的员工培训。假设共有需要培训的需求(如技术类、管理类)为 6 种，每种需求的最低培训人数为 $a_i, i=1,\cdots,6$，可供选择的培训方式(如内部自行培训、外部与高校合作培训)有 5 种，每种的最高培训人数为 $b_j, j=1,\cdots,5$。又设若选择了第 1 种培训方式，则第 3 种培训方式也要选择。记 x_{ij} 为第 i 种需求由第 j 种方式培训的人员数量，Z 为培训总费用。费用的构成包括固定费用和可变费用，第 j 种方式的固定培训费用为 h_j(与人数无关)，与人数 x_{ij} 相应的可变费用为 C_{ij}(表示用第 j 种方式培训第 i 种需求类型的单位费用)。如果以成本费用为优化目标，请建立该培训问题的结构优化模型。

6.6 为了提高校园的安全性，某大学的保安部门决定在校园内部的几个位置安装紧急报警电话。校园的主要街道示意图如图 6-9，其中①～⑧表示道路交叉口，$A\sim K$ 表示街道。现需决定在哪些地方安装，可使每条街道都有报警电话，并且总电话数目最少？请建立本问题的数学规划模型。

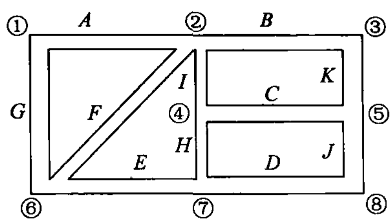


图 6-9

6.7 在有互相排斥的约束条件的问题中，如果约束条件是(≤)型的，我们用加以 y_iM 项(y_i 是 0-1 变量，M 是很大的常数)的方法统一在一个问题中。

如果约束条件是(≥)型的，我们将怎样利用 y_i 和 M 呢？

6.8 解 0-1 规划

(1) $\min z=4x_1+3x_2+2x_3$

$$\begin{cases}2x_1-5x_2+3x_3\leqslant 4\\4x_1+x_2+3x_3\geqslant 3\\x_2+x_3\geqslant 1\\x_1,x_2,x_3=0\text{ 或 }1\end{cases}$$

(2) $\min z=2x_1+5x_2+3x_3+4x_4$

$$\begin{cases}-4x_1+x_2+x_3+x_4\geqslant 0\\-2x_1+4x_2+2x_3+4x_4\geqslant 4\\x_1+x_2-x_3+x_4\geqslant 1\\x_1,x_2,x_3,x_4=0\text{ 或 }1\end{cases}$$

6.9 有 4 个工人，要指派他们分别完成 4 种工作，每人做各种工作所消耗的时间如表 6-10 所示，问指派哪个人去完成哪种工作，可使总的消耗时间为最小？

表 6-10

工人	工种			
	A	*B*	*C*	*D*
甲	15	18	21	24
乙	19	23	22	18
丙	26	17	16	19
丁	19	21	23	17

参考资料

[1] A. Kaufmann. Integer and mixed programming. *Theory and Application*. Academic press. INC (London)LTD.

[2] 孙小玲，李端. 整数规划，运筹与管理科学丛书 10. 科学出版社，2010.

Part 4 第4篇

非线性规划

由前面几章知道，在科学管理和其他领域中，很多实际问题可以归结为线性规划问题，其目标函数和约束条件都是自变量的一次函数。但是，还有另外一些问题，其目标函数和(或)约束条件很难用线性函数表达。如果目标函数或约束条件中含有**非线性函数**，就称这种规划问题为**非线性规划**问题。解这种问题要用非线性规划的方法。由于很多实际问题要求进一步精确化以及电子计算机的发展，使非线性规划在近几十年来得以长驱进展。目前，它已成为运筹学的重要分支之一，并在最优设计、管理科学、系统控制等许多领域得到越来越广泛的应用。

一般来说，由于非线性函数的复杂性，解非线性规划问题要比解线性规划问题困难得多。而且，也不像线性规划有单纯形法等通用方法，非线性规划目前还没有适于解决各种问题的一般算法，各个方法都有自己特定的适用范围。这是需要人们更深入地进行研究的一个领域。

在以下两章中，除了简要地介绍非线性规划的基本概念和一维搜索法之外，着重说明无约束极值问题和约束极值问题的主要解法。为了叙述方便，我们常用大写字母代表 n 维欧氏空间中的向量(点)，而以相应的小写字母代表该向量的分量(点的坐标)。此外，在这一部分所用到的向量，均规定为列向量。

CHAPTER 7

第7章*

无约束问题

7.1 基本概念

7.1.1 引言

1. 问题的提出

让我们先看两个例子。

例1 某公司经营两种产品,第一种产品每件售价 30 元,第二种产品每件售价 450 元。根据统计,售出一件第一种产品所需要的服务时间平均是 0.5 小时,第二种产品是 $(2+0.25x_2)$小时,其中 x_2 是第二种产品的售出数量。已知该公司在这段时间内的总服务时间为 800 小时,试决定使其营业额最大的营业计划。

下面我们来分析这个例子,并为其建立数学模型。

设该公司计划经营第一种产品 x_1 件,第二种产品 x_2 件。根据题意,其营业额为

$$f(X)=30x_1+450x_2$$

由于服务时间的限制,该计划必须满足

$$0.5x_1+(2+0.25x_2)x_2\leqslant 800$$

此外,这个问题还应满足

$$x_1\geqslant 0,\quad x_2\geqslant 0$$

如此,得到这个问题的数学模型如下

$$\begin{cases}\max\ f(X)=30x_1+450x_2\\ \quad 0.5x_1+2x_2+0.25x_2^2\leqslant 800\\ \quad x_1\geqslant 0,\quad x_2\geqslant 0\end{cases}$$

例2 为了进行多属性问题(假设有 n 个属性)的综合评价,就需要知道每个属性的相对重要性,即确定它们的权重。为此将各属性的重要性(对评价者或决策者而言)进行两两比较,从而得出如下判断矩阵

$$\boldsymbol{J}=\begin{bmatrix}a_{11} & \cdots & a_{1n}\\ \vdots & \ddots & \vdots\\ a_{n1} & \cdots & a_{nn}\end{bmatrix}$$

其中元素 a_{ij} 是第 i 个属性的重要性与第 j 个属性的重要性之比。

现需从判断矩阵求出各属性的权重 $w_i(i=1,2,\cdots,n)$。为了使求出的权向量

$$W=(w_1,w_2,\cdots,w_n)^{\mathrm{T}}$$

在最小二乘意义上能最好的反映判断矩阵的估计，由 $a_{ij}\approx w_i/w_j$，可得

$$\begin{cases}\min \sum_{i=1}^{n}\sum_{j=1}^{n}(a_{ij}w_j-w_i)^2\\ \sum_{i=1}^{n}w_i=1\end{cases}$$

例 1 的目标函数为自变量的线性函数，但其第一个约束条件却是自变量的二次函数，因而它是非线性规划问题。例 2 的目标函数是自变量的非线性函数，所以它也是非线性规划问题。事实上，非线性规划问题广泛存在于管理和决策问题之中。

2. 非线性规划问题的数学模型

非线性规划的数学模型常表示成以下形式

$$\begin{cases}\min f(\boldsymbol{X}) & (7\text{-}1)\\ h_i(\boldsymbol{X})=0,i=1,2,\cdots,m & (7\text{-}2)\\ g_j(\boldsymbol{X})\geqslant 0,j=1,2,\cdots,l & (7\text{-}3)\end{cases}$$

其中自变量 $\boldsymbol{X}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$ 是 n 维欧氏空间 E^n 中的向量(点)；$f(\boldsymbol{X})$为目标函数，$h_i(\boldsymbol{X})=0$ 和 $g_j(\boldsymbol{X})\geqslant 0$ 为约束条件。

由于 $\max f(\boldsymbol{X})=-\min[-f(\boldsymbol{X})]$，当需使目标函数极大化时，只需使其负值极小化即可。因而仅考虑目标函数极小化，这无损于一般性。

若某约束条件是“$\leqslant$”不等式时，仅需用“-1”乘该约束的两端，即可将这个约束变为“$\geqslant$”的形式。

由于等式约束

$$h_i(\boldsymbol{X})=0$$

等价于下述两个不等式约束

$$\begin{cases}h_i(\boldsymbol{X})\geqslant 0\\ -h_i(\boldsymbol{X})\geqslant 0\end{cases}$$

因而，也可将非线性规划的数学模型写成以下形式

$$\begin{cases}\min f(\boldsymbol{X}) & (7\text{-}4)\\ g_j(\boldsymbol{X})\geqslant 0,\quad j=1,2,\cdots,l & (7\text{-}5)\end{cases}$$

3. 非线性规划问题的图示

图示法可以给人以直观概念，当只有两个自变量时，非线性规划问题也可像线性规划那样用图示法来表示(如图 7-1 所示)。

考虑非线性规划问题

$$
\begin{cases}
\min f(\boldsymbol{X})=(x_1-2)^2+(x_2-2)^2 & (7\text{-}6)\\
h(\boldsymbol{X})=x_1+x_2-6=0 & (7\text{-}7)
\end{cases}
$$

若令其目标函数

$$f(\boldsymbol{X})=c \tag{7-8}$$

其中，c 为某一常数，则式(7-8)代表目标函数值等于 c 的点的集合，它一般为一条曲线或一张曲面，通常称其为**等值线**或**等值面**。对于这个例子来说，若令目标函数式(7-6)分别等于 2 和 4，就得到相应的两条圆形等值线(见图 7-1)。由图可见，等值线 $f(\boldsymbol{X})=2$ 和约束条件直线 AB 相切，切点 D 即为此问题的**最优解**：$x_1^*=x_2^*=3$，其目标函数值 $f(\boldsymbol{X}^*)=2$。

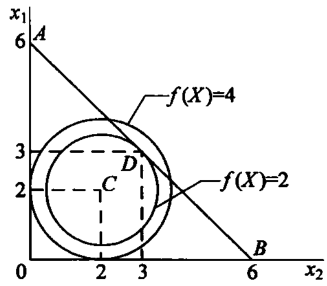


图 7-1

在这个例子中，约束条件式(7-7)对最优解是有影响的。

现若以

$$h(\boldsymbol{X})=x_1+x_2-6\leqslant 0 \tag{7-9}$$

代替约束条件式(7-7)，则非线性规划问题式(7-6)、式(7-9)的最优解是 $x_1=x_2=2$，即图 7-1 中的 C 点(这时 $f(\boldsymbol{X})=0$)。由于最优点位于可行域的内部，故对这个问题的最优解来说，约束式(7-9)事实上是不起作用的。在求这个问题的最优解时，可不考虑约束条件式(7-9)，就相当于没有这个约束一样。

由第 2 章知道，如果线性规划问题的最优解存在，其最优解只能在其可行域的边界上达到(特别是在可行域的顶点上达到)；而非线性规划问题的最优解(如果最优解存在)则可能在其可行域中的任意一点达到。

7.1.2 极值问题

在高等数学课程中，已学过一元函数和多元函数的极值问题，现仅扼要说明如下。

1. 局部极值和全局极值

由于线性规划的目标函数为线性函数，可行域为凸集，因而求出的最优解就是在整个可行域上的全局最优解。非线性规划却不然，有时求出的某个解虽是一部分可行域上的极值点，但却并不一定是整个可行域上的全局最优解。

设 $f(\boldsymbol{X})$ 为定义在 n 维欧氏空间 E^n 的某一区域 R 上的 n 元实函数，其中 $\boldsymbol{X}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$。对于 $\boldsymbol{X}^*\in R$，如果存在某个 $\varepsilon>0$，使所有与 $\boldsymbol{X}^*$ 的距离小于 ε 的 $\boldsymbol{X}\in R$(即 $\boldsymbol{X}\in R$ 且 $\|\boldsymbol{X}-\boldsymbol{X}^*\|<\varepsilon$)均满足不等式 $f(\boldsymbol{X})\geqslant f(\boldsymbol{X}^*)$，则称 $\boldsymbol{X}^*$ 为 $f(\boldsymbol{X})$ 在 R 上的**局部极小点**(或相对极小点)，$f(\boldsymbol{X}^*)$ 为**局部极小值**。若对于所有 $\boldsymbol{X}\neq\boldsymbol{X}^*$ 且与 $\boldsymbol{X}^*$ 的距离小于 ε 的 $\boldsymbol{X}\in R$，$f(\boldsymbol{X})>f(\boldsymbol{X}^*)$，则称 $\boldsymbol{X}^*$ 为 $f(\boldsymbol{X})$ 在 R 上的**严格局格极小点**，$f(\boldsymbol{X}^*)$ 为**严格**

局部极小值。

若点 $\boldsymbol{X}^* \in R$，而对于所有 $\boldsymbol{X} \in R$ 都有 $f(\boldsymbol{X}) \geqslant f(\boldsymbol{X}^*)$，则称 $\boldsymbol{X}^*$ 为 $f(\boldsymbol{X})$ 在 R 上的**全局极小点**，$f(\boldsymbol{X}^*)$ 为**全局极小值**。若对于所有 $\boldsymbol{X} \in R$ 且 $\boldsymbol{X} \neq \boldsymbol{X}^*$，都有 $f(\boldsymbol{X}) > f(\boldsymbol{X}^*)$，则称 $\boldsymbol{X}^*$ 为 $f(\boldsymbol{X})$ 在 R 上的**严格全局极小点**，$f(\boldsymbol{X}^*)$ 为**严格全局极小值**。

如将上述不等式反向，即可得到相应的极大点和极大值的定义。

下面仅就极小点及极小值加以说明，而且主要研究局部极小。

2. 极值点存在的条件

现说明极值点存在的必要条件和充分条件。

定理 1 （必要条件）

设 R 是 n 维欧氏空间 E^n 上的某一开集，$f(\boldsymbol{X})$ 在 R 上有一阶连续偏导数，且在点 $\boldsymbol{X}^* \in R$ 取得局部极值，则必有

$$\frac{\partial f(\boldsymbol{X}^*)}{\partial x_1}=\frac{\partial f(\boldsymbol{X}^*)}{\partial x_2}=\cdots=\frac{\partial f(\boldsymbol{X}^*)}{\partial x_n}=0 \tag{7-10}$$

或

$$\nabla f(\boldsymbol{X}^*)=0 \tag{7-11}$$

上式中

$$\nabla f(\boldsymbol{X}^*)=\left(\frac{\partial f(\boldsymbol{X}^*)}{\partial x_1},\frac{\partial f(\boldsymbol{X}^*)}{\partial x_2},\cdots,\frac{\partial f(\boldsymbol{X}^*)}{\partial x_n}\right)^{\mathrm{T}} \tag{7-12}$$

为函数 $f(\boldsymbol{X})$ 在点 $\boldsymbol{X}^*$ 处的**梯度**。

由数学分析知道，$\nabla f(\boldsymbol{X})$ 的方向为 $f(\boldsymbol{X})$ 的等值面（等值线）的法线（在点 $\boldsymbol{X}$ 处）方向，沿这个方向函数值增加最快。

满足式(7-10)或式(7-11)的点称为**平稳点**或**驻点**，在区域内部，极值点必为平稳点，但平稳点不一定是极值点。

定理 2 （充分条件）

设 R 是 n 维欧氏空间 E^n 上的某一开集，$f(\boldsymbol{X})$ 在 R 上具有二阶连续偏导数，$\boldsymbol{X}^* \in R$，若 $\nabla f(\boldsymbol{X}^*)=0$，且对任何非零向量 $\boldsymbol{Z} \in E^n$ 有

$$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{X}^*)\boldsymbol{Z}>0 \tag{7-13}$$

则 $\boldsymbol{X}^*$ 为 $f(\boldsymbol{X})$ 的严格局部极小点。

此处 $\boldsymbol{H}(\boldsymbol{X}^*)$ 为 $f(\boldsymbol{X})$ 在点 $\boldsymbol{X}^*$ 处的**海赛(Hesse)矩阵**

$$\boldsymbol{H}(\boldsymbol{X}^*)=\begin{bmatrix} \frac{\partial^2 f(\boldsymbol{X}^*)}{\partial x_1^2} & \frac{\partial^2 f(\boldsymbol{X}^*)}{\partial x_1\partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{X}^*)}{\partial x_1\partial x_n} \\ \frac{\partial^2 f(\boldsymbol{X}^*)}{\partial x_2\partial x_1} & \frac{\partial^2 f(\boldsymbol{X}^*)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\boldsymbol{X}^*)}{\partial x_2\partial x_n} \\ & & \vdots & \\ \frac{\partial^2 f(\boldsymbol{X}^*)}{\partial x_n\partial x_1} & \frac{\partial^2 f(\boldsymbol{X}^*)}{\partial x_n\partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{X}^*)}{\partial x_n^2} \end{bmatrix} \tag{7-14}$$

证明从略。

需要指出，定理 2 中的充分条件式(7-13)并不是必要的。可以举出这样的例子：X^* 是 $f(X)$的极小点，但却不满足条件式(7-13)。例如，$f(x)=x^4$，它的极小点是 $x^*=0$，但 $f''(x^*)=0$，这不满足式(7-13)。

二次型是 $X=(x_1,x_2,\cdots,x_n)^T$ 的二次齐次函数，它在研究非线性最优化中具有重要作用。现考虑**二次型** Z^THZ。若对于任意 $Z\neq 0$(即 Z 的元素不全为零)，二次型 Z^THZ 的值总是正的，即 $Z^THZ>0$，则称该二次型是**正定**的；若对于任意 $Z\neq 0$ 总有 $Z^THZ\geqslant 0$，则称其为**半正定**；若对于任意 $Z\neq 0$ 总有 $Z^THZ<0$，则称其为**负定**；若对于任意 $Z\neq 0$ 总有 $Z^THZ\leqslant 0$，则称其为**半负定**。如果对某些 $Z\neq 0$，$Z^THZ>0$，而对另一些 $Z\neq 0$，$Z^THZ<0$，即它既非正定，也非负定，则称其为**不定的**。由线性代数学知道，二次型 Z^THZ 为正定的充要条件，是它的矩阵 H 的左上角各阶主子式都大于零；而它为负定的充要条件，是它的矩阵 H 的左上角各阶主子式依次负正相间。

现以 h_{ij} 表示矩阵 H 的元素，上述条件为，当二次型正定时

$$h_{11}>0;\qquad \begin{vmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{vmatrix}>0;\qquad \cdots;\qquad \begin{vmatrix} h_{11} & \cdots & h_{1n} \\ & \cdots & \\ h_{n1} & \cdots & h_{nn} \end{vmatrix}>0$$

当二次型负定时

$$h_{11}<0;\qquad \begin{vmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{vmatrix}>0;$$

$$\begin{vmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{vmatrix}<0;\qquad \cdots;\qquad (-1)^n\begin{vmatrix} h_{11} & \cdots & h_{1n} \\ & \cdots & \\ h_{n1} & \cdots & h_{nn} \end{vmatrix}>0$$

二次型 Z^THZ 为正定、负定或不定时，其对称矩阵 H 分别称为正定的、负定的或不定的。定理 2 中的条件式(7-13)，就等于说其海赛矩阵在 X^* 处正定。

7.1.3 凸函数和凹函数

凸集、凸函数以及凸函数的极值的性质，是研究非线性规划问题所不可缺少的内容。凸集的概念在讲线性规划时已作过说明，因而这里简要说明凸函数的有关问题。

1. 什么是凸函数和凹函数

设 $f(X)$为定义在 n 维欧氏空间 E^n 中某个凸集 R 上的函数，若对任何实数 $\alpha(0<\alpha<1)$ 以及 R 中的任意两点 $X^{(1)}$ 和 $X^{(2)}$，恒有

$$f(\alpha X^{(1)}+(1-\alpha)X^{(2)})\leqslant \alpha f(X^{(1)})+(1-\alpha)f(X^{(2)}) \tag{7-15}$$

则称 $f(X)$为定义在 R 上的**凸函数**。

若对任意 $\alpha(0<\alpha<1)$和 $X^{(1)}\neq X^{(2)}\in R$ 恒有

$$f(\alpha \boldsymbol{X}^{(1)}+(1-\alpha)\boldsymbol{X}^{(2)})<\alpha f(\boldsymbol{X}^{(1)})+(1-\alpha)f(\boldsymbol{X}^{(2)}) \tag{7-16}$$

则称 $f(\boldsymbol{X})$ 为定义在 R 上的**严格凸函数**。

将式(7-15)和式(7-16)中的不等号反向，即可得到**凹函数**和**严格凹函数**的定义。显然，若函数 $f(\boldsymbol{X})$ 是凸函数(严格凸函数)，则 $-f(\boldsymbol{X})$ 一定是凹函数(严格凹函数)。

凸函数和凹函数的几何意义十分明显，若函数图形上任两点的连线处都不在这个函数图形的下方，它当然是下凸的(见图 7-2(a))。凹函数则是下凹的(上凸的)(见图 7-2(b))。线性函数既可看做凸函数，也可看做凹函数。

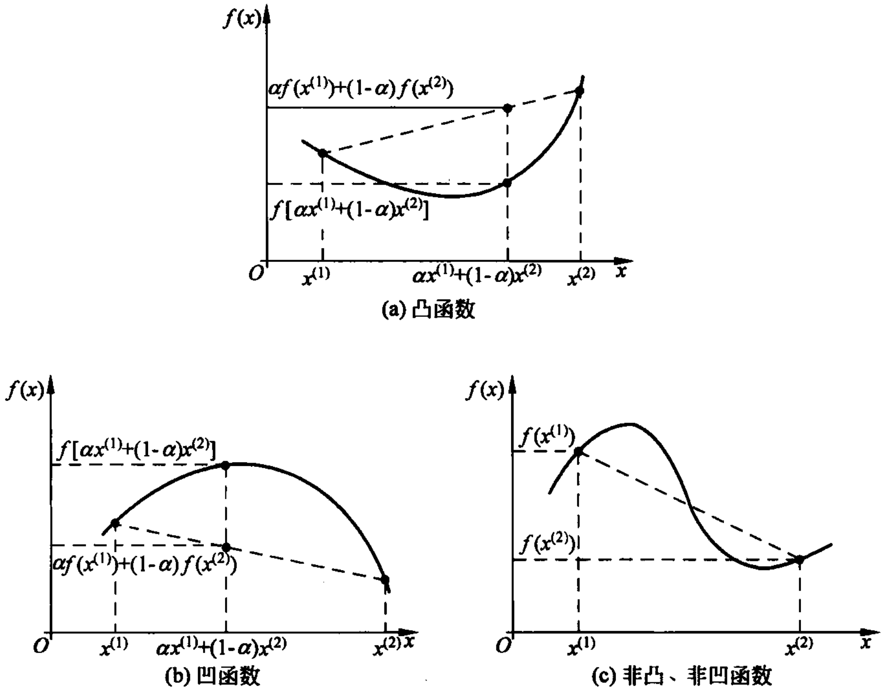


图 7-2

2. 凸函数的性质

凸函数具有如下性质：

性质 1 设 $f(\boldsymbol{X})$ 为定义在凸集 R 上的凸函数，则对任意实数 $\beta \geqslant 0$，函数 $\beta f(\boldsymbol{X})$ 也是定义在 R 上的凸函数。

性质 2 设 $f_1(\boldsymbol{X})$ 和 $f_2(\boldsymbol{X})$ 为定义在凸集 R 上的两个凸函数，则其和 $f(\boldsymbol{X})=f_1(\boldsymbol{X})+f_2(\boldsymbol{X})$ 仍为定义在 R 上的凸函数。

因为 $f_1(\boldsymbol{X})$ 和 $f_2(\boldsymbol{X})$ 都是定义在 R 上的凸函数，故对 R 上的任两点 $\boldsymbol{X}^{(1)}$ 和 $\boldsymbol{X}^{(2)}$ 以及任意实数 $\alpha(0<\alpha<1)$ 恒有

$$f_1(\alpha \boldsymbol{X}^{(1)}+(1-\alpha)\boldsymbol{X}^{(2)}) \leqslant \alpha f_1(\boldsymbol{X}^{(1)})+(1-\alpha)f_1(\boldsymbol{X}^{(2)})$$

$$f_2(\alpha \boldsymbol{X}^{(1)}+(1-\alpha)\boldsymbol{X}^{(2)})\leqslant \alpha f_2(\boldsymbol{X}^{(1)})+(1-\alpha)f_2(\boldsymbol{X}^{(2)})$$

将上述两式两端分别相加得

$$f(\alpha \boldsymbol{X}^{(1)}+(1-\alpha)\boldsymbol{X}^{(2)})\leqslant \alpha f(\boldsymbol{X}^{(1)})+(1-\alpha)f(\boldsymbol{X}^{(2)})$$

故 $f(\boldsymbol{X})$ 也是 R 上的凸函数。

由以上两个性质立刻推得：有限个凸函数的非负线性组合

$$\beta_1 f_1(\boldsymbol{X})+\beta_2 f_2(\boldsymbol{X})+\cdots+\beta_m f_m(\boldsymbol{X})$$

$$\beta_i\geqslant 0,\quad i=1,2,\cdots,m$$

仍为凸函数。

性质 3 设 $f(\boldsymbol{X})$ 为定义在凸集 R 上的凸函数，则对任一实数 β，集合

$$S_\beta=\{\boldsymbol{X}\mid \boldsymbol{X}\in R, f(\boldsymbol{X})\leqslant\beta\} \tag{7-17}$$

是凸集（S_β 称为水平集）。

证明 任取 $\boldsymbol{X}^{(1)}\in S_\beta$ 和 $\boldsymbol{X}^{(2)}\in S_\beta$，则有 $f(\boldsymbol{X}^{(1)})\leqslant\beta, f(\boldsymbol{X}^{(2)})\leqslant\beta$。

由于 R 为凸集，故对任意实数 $\alpha(0<\alpha<1)$，$\alpha\boldsymbol{X}^{(1)}+(1-\alpha)\boldsymbol{X}^{(2)}\in R$，又因 $f(\boldsymbol{X})$ 为凸函数，故

$$f(\alpha \boldsymbol{X}^{(1)}+(1-\alpha)\boldsymbol{X}^{(2)})\leqslant \alpha f(\boldsymbol{X}^{(1)})+(1-\alpha)f(\boldsymbol{X}^{(2)})\leqslant\beta$$

这就表明点 $\alpha\boldsymbol{X}^{(1)}+(1-\alpha)\boldsymbol{X}^{(2)}\in S_\beta$，于是，$S_\beta$ 为凸集。

3. 函数凸性的判定

现在来研究怎样判断一个函数是凸函数，当然可以直接依据定义去判别。对于可微凸函数，也可利用下述两个判别定理。

定理 3 （一阶条件）

设 R 为 n 维欧氏空间 E^n 上的开凸集，$f(\boldsymbol{X})$ 在 R 上具有一阶连续偏导数，则 $f(\boldsymbol{X})$ 为 R 上的凸函数的充要条件是，对任意两个不同点 $\boldsymbol{X}^{(1)}\in R$ 和 $\boldsymbol{X}^{(2)}\in R$，恒有

$$f(\boldsymbol{X}^{(2)})\geqslant f(\boldsymbol{X}^{(1)})+\nabla f(\boldsymbol{X}^{(1)})^{\mathrm{T}}(\boldsymbol{X}^{(2)}-\boldsymbol{X}^{(1)}) \tag{7-18}$$

证明 必要性：

设 $f(\boldsymbol{X})$ 为 R 上的凸函数，则对任何 $\alpha(0<\alpha<1)$ 有

$$f(\alpha \boldsymbol{X}^{(2)}+(1-\alpha)\boldsymbol{X}^{(1)})\leqslant \alpha f(\boldsymbol{X}^{(2)})+(1-\alpha)f(\boldsymbol{X}^{(1)})$$

于是

$$\frac{f(\boldsymbol{X}^{(1)}+\alpha(\boldsymbol{X}^{(2)}-\boldsymbol{X}^{(1)}))-f(\boldsymbol{X}^{(1)})}{\alpha}\leqslant f(\boldsymbol{X}^{(2)})-f(\boldsymbol{X}^{(1)})$$

令 $\alpha\to+0$，上式左端的极限为

$$\nabla f(\boldsymbol{X}^{(1)})^{\mathrm{T}}(\boldsymbol{X}^{(2)}-\boldsymbol{X}^{(1)})$$

即

$$f(\boldsymbol{X}^{(2)})\geqslant f(\boldsymbol{X}^{(1)})+\nabla f(\boldsymbol{X}^{(1)})^{\mathrm{T}}(\boldsymbol{X}^{(2)}-\boldsymbol{X}^{(1)})$$

充分性：

任取 $\boldsymbol{X}^{(1)}\in R$ 及 $\boldsymbol{X}^{(2)}\in R$，现令

$$\boldsymbol{X}=\alpha\boldsymbol{X}^{(1)}+(1-\alpha)\boldsymbol{X}^{(2)},\quad 0<\alpha<1$$

分别以 $X^{(1)}$ 和 $X^{(2)}$ 为式(7-18)中的 $X^{(2)}$，以 X 为式(7-18)中的 $X^{(1)}$，则

$$f(X^{(1)})\geqslant f(X)+\nabla f(X)^{\mathrm{T}}(X^{(1)}-X)$$

$$f(X^{(2)})\geqslant f(X)+\nabla f(X)^{\mathrm{T}}(X^{(2)}-X)$$

用 α 乘上面的第一式，用 $(1-\alpha)$ 乘上面的第二式，然后两端相加

$$\alpha f(X^{(1)})+(1-\alpha)f(X^{(2)})\geqslant f(X)+\nabla f(X)^{\mathrm{T}}\times[\alpha X^{(1)}-\alpha X+(1-\alpha)(X^{(2)}-X)]$$
$$=f(X)=f(\alpha X^{(1)}+(1-\alpha)X^{(2)})$$

从而可知 $f(X)$ 为 R 上的凸函数。

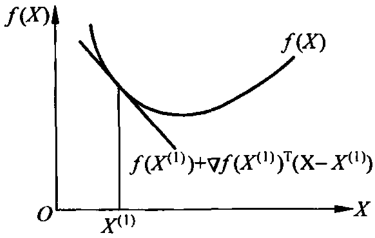


图 7-3

若式(7-18)为严格不等式，它就是严格凸函数的充要条件。

凸函数的定义式(7-15)，本质上是说凸函数上两点间的线性插值不低于这个函数的值；而定理 3 则是说，基于某点导数的线性近似不高于这个函数的值(图 7-3)。

定理 4 (二阶条件)

设 R 为 n 维欧氏空间 E^n 上的某一开凸集，$f(X)$ 在 R 上具有二阶连续偏导数，则 $f(X)$ 为 R 上的凸函数的充要条件是：$f(X)$ 的海赛矩阵 $H(X)$ 在 R 上处处半正定。

证明 先证必要性。

设 $f(X)$ 为 R 上的凸函数。任取 $X\in R$ 和 $Z\in E^n$，现证

$$Z^{\mathrm{T}}H(X)Z\geqslant 0$$

因 R 为开集，故存在 $\bar{\alpha}>0$，使当 $\alpha=[-\bar{\alpha},\bar{\alpha}]$ 时，有 $X+\alpha Z\in R$。由定理 3 可得

$$f(X+\alpha Z)\geqslant f(X)+\alpha\nabla f(X)^{\mathrm{T}}Z$$

再由泰勒公式

$$f(X+\alpha Z)=f(X)+\alpha\nabla f(X)^{\mathrm{T}}Z+\frac{1}{2}\alpha^2Z^{\mathrm{T}}H(X)Z+o(\alpha^2)$$

其中

$$\lim_{\alpha\to 0}\frac{o(\alpha^2)}{\alpha^2}=0$$

由以上两式得

$$\frac{1}{2}\alpha^2Z^{\mathrm{T}}H(X)Z+o(\alpha^2)\geqslant 0$$

从而

$$\frac{1}{2}Z^{\mathrm{T}}H(X)Z+\frac{o(\alpha^2)}{\alpha^2}\geqslant 0$$

令 $\alpha\to 0$，则得

$$Z^{\mathrm{T}}H(X)Z\geqslant 0$$

即 $\boldsymbol{H}(\boldsymbol{X})$ 为半正定矩阵。

下面证明充分性。设对任意 $\boldsymbol{X}\in R$，$\boldsymbol{H}(\boldsymbol{X})$ 为半正定矩阵，任取 $\bar{\boldsymbol{X}}\in R$，由泰勒公式，有

$$f(\boldsymbol{X})=f(\bar{\boldsymbol{X}})+\nabla f(\bar{\boldsymbol{X}})^{\mathrm{T}}(\boldsymbol{X}-\bar{\boldsymbol{X}})+\frac{1}{2}(\boldsymbol{X}-\bar{\boldsymbol{X}})^{\mathrm{T}}H(\bar{\boldsymbol{X}}+\lambda(X-\bar{\boldsymbol{X}}))(X-\bar{\boldsymbol{X}})$$

其中 $\lambda\in(0,1)$。

因 R 为凸集，$\bar{\boldsymbol{X}}+\lambda(X-\bar{\boldsymbol{X}})\in R$。再由假设知 $\boldsymbol{H}(\bar{\boldsymbol{X}}+\lambda(\boldsymbol{X}-\bar{\boldsymbol{X}}))$ 为半正定，从而

$$f(\boldsymbol{X})\geqslant f(\bar{\boldsymbol{X}})+\nabla f(\bar{\boldsymbol{X}})^{\mathrm{T}}(\boldsymbol{X}-\bar{\boldsymbol{X}})$$

由定理 3，$f(\boldsymbol{X})$ 为 R 上的凸函数。

若对一切 $\boldsymbol{X}\in R$，$f(\boldsymbol{X})$ 的海赛矩阵都是正定的，则 $f(\boldsymbol{X})$ 是 R 上的严格凸函数。

对于凹函数可以得到和上述类似的结果。

例 3 试证明 $f(\boldsymbol{X})=-x_1^2-x_2^2$ 为凹函数。

证 首先由定义证明 $f_1(x_1)=-x_1^2$ 为凹函数。

任意指定两点 a_1 和 a_2，看下述各式是否成立？

$$-[\alpha a_1+(1-\alpha)a_2]^2\geqslant\alpha(-a_1^2)+(1-\alpha)(-a_2^2)$$

或

$$a_1^2(\alpha-\alpha^2)-2a_1a_2(\alpha-\alpha^2)+a_2^2(\alpha-\alpha^2)\geqslant 0$$

或

$$(\alpha-\alpha^2)(a_1-a_2)^2\geqslant 0$$

由于 $0<\alpha<1$，故 $\alpha-\alpha^2>0$。显然，不管 a_1 和 a_2 取什么值，总有 $(\alpha-\alpha^2)(a_1-a_2)^2\geqslant 0$ 成立，从而证明 $f_1(x_1)=-x_1^2$ 为凹函数。用同样的方法可以证明 $f_2(x_2)=-x_2^2$ 也是凹函数。根据性质 2，$f(\boldsymbol{X})=-x_1^2-x_2^2$ 为凹函数。

再用定理 3 证明。任意选取第一点 $\boldsymbol{X}^{(1)}=(a_1,b_1)^{\mathrm{T}}$，第二点 $\boldsymbol{X}^{(2)}=(a_2,b_2)^{\mathrm{T}}$。如此，

$$f(\boldsymbol{X}^{(1)})=-a_1^2-b_1^2\qquad f(\boldsymbol{X}^{(2)})=-a_2^2-b_2^2$$

$$\nabla f(\boldsymbol{X})=(-2x_1,-2x_2)^{\mathrm{T}}$$

$$\nabla f(\boldsymbol{X}^{(1)})=(-2a_1,-2b_1)^{\mathrm{T}}$$

现看下述各式是否成立？

$$-a_2^2-b_2^2\leqslant -a_1^2-b_1^2+(-2a_1\quad -2b_1)\begin{pmatrix}a_2-a_1\\b_2-b_1\end{pmatrix}$$

或

$$-a_2^2-b_2^2\leqslant -a_1^2-b_1^2-2a_1(a_2-a_1)-2b_1(b_2-b_1)$$

或

$$-(a_2^2-2a_1a_2+a_1^2)-(b_2^2-2b_1b_2+b_1^2)\leqslant 0$$

或

$$-(a_2-a_1)^2-(b_2-b_1)^2\leqslant 0$$

不管 a_1、a_2、b_1 和 b_2 取什么值，上式均成立，从而得证。

下面用定理 4 证明。由于

$$\frac{\partial f(\boldsymbol{X})}{\partial x_1}=-2x_1 \qquad \frac{\partial f(\boldsymbol{X})}{\partial x_2}=-2x_2$$

$$\frac{\partial^2 f(\boldsymbol{X})}{\partial x_1^2}=-2<0 \qquad \frac{\partial^2 f(\boldsymbol{X})}{\partial x_2^2}=-2$$

$$\frac{\partial^2 f(\boldsymbol{X})}{\partial x_1 \partial x_2}=\frac{\partial^2 f(\boldsymbol{X})}{\partial x_2 \partial x_1}=0$$

$$|\boldsymbol{H}|=\begin{vmatrix} -2 & 0 \\ 0 & -2 \end{vmatrix}=4>0$$

其海赛矩阵处处负定，故 $f(\boldsymbol{X})$ 为(严格)凹函数。

4. 凸函数的极值

前已指出，函数的局部极小值并不一定等于它的最小值，前者只不过反映了函数的局部性质。而最优化的目的，往往是要求函数在整个域中的最小值(或最大值)和最小点(或最大点)。为此，必须将所得的全部极小值进行比较(有时尚需考虑边界值)，以便从中选出最小者。然而，对于定义在凸集上的凸函数来说，则用不着进行这种麻烦的工作，它的极小值就等于其最小值。

定理 5 若 $f(\boldsymbol{X})$ 为定义在凸集 R 上的凸函数，则它的任一极小点就是它在 R 上的最小点(全局极小点)，而且它的极小点形成一个凸集。

证明 设 $\boldsymbol{X}^*$ 是一个局部极小点，则对于充分小的邻域 $N_\delta(\boldsymbol{X}^*)$ 中的一切 $\boldsymbol{X}$，均有

$$f(\boldsymbol{X})\geqslant f(\boldsymbol{X}^*)$$

令 $\boldsymbol{Y}$ 是 R 中的任一点，对于充分小的 λ，$0<\lambda<1$，就有

$$((1-\lambda)\boldsymbol{X}^*+\lambda\boldsymbol{Y})\in N_\delta(\boldsymbol{X}^*)$$

从而

$$f((1-\lambda)\boldsymbol{X}^*+\lambda\boldsymbol{Y})\geqslant f(\boldsymbol{X}^*)$$

由于 $f(\boldsymbol{X})$ 为凸函数，故

$$(1-\lambda)f(\boldsymbol{X}^*)+\lambda f(\boldsymbol{Y})\geqslant f((1-\lambda)\boldsymbol{X}^*+\lambda\boldsymbol{Y})$$

将上述两个不等式相加，移项后除以 λ，得到

$$f(\boldsymbol{Y})\geqslant f(\boldsymbol{X}^*)$$

这就是说，$\boldsymbol{X}^*$ 是全局极小点。

由性质 3，所有极小点的集合形成一个凸集。

定理 6 设 $f(\boldsymbol{X})$ 是定义在凸集 R 上的可微凸函数，若存在点 $\boldsymbol{X}^*\in R$，使得对于所有的 $\boldsymbol{X}\in R$ 有

$$\nabla f(\boldsymbol{X}^*)^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{X}^*)\geqslant 0 \tag{7-19}$$

则 $\boldsymbol{X}^*$ 是 $f(\boldsymbol{X})$ 在 R 上的最小点(全局极小点)。

证明 由定理 3

$$f(\boldsymbol{X}) \geqslant f(\boldsymbol{X}^*) + \nabla f(\boldsymbol{X}^*)^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{X}^*)$$

如此，对所有 $\boldsymbol{X} \in R$ 有

$$f(\boldsymbol{X}) \geqslant f(\boldsymbol{X}^*)$$

一种极为重要的情形是，当点 $\boldsymbol{X}^*$ 是 R 的内点时，这时式(7-19)对任意 $\boldsymbol{X}-\boldsymbol{X}^*$ 都成立，这就意味着可将式(7-19)改为

$$\nabla f(\boldsymbol{X}^*) = 0$$

以上两个定理说明，定义在凸集上的凸函数的平稳点，就是其全局极小点。全局极小点并不一定是唯一的，但若为严格凸函数，则其全局极小点就是唯一的了。

7.1.4 凸规则

现在再回到非线性规划式(7-1)～式(7-3)。和线性规划类似，把满足约束条件式(7-2)和式(7-3)的点称做**可行点**(**可行解**)，所有可行点的集合称做**可行域**。若某个可行解使目标函数式(7-1)为最小，就称它为**最优解**。

考虑非线性规划

$$\begin{cases} \min\limits_{X \in R} f(\boldsymbol{X}) \\ R = \{\boldsymbol{X} \mid g_j(\boldsymbol{X}) \geqslant 0, \quad j = 1, 2, \cdots, l\} \end{cases}$$

假定其中 $f(\boldsymbol{X})$ 为凸函数，$g_j(\boldsymbol{X})(j=1,2,\cdots,l)$ 为凹函数(或者说 $-g_j(\boldsymbol{X})$ 为凸函数)，这样的非线性规划称为**凸规划**。可以证明，上述凸规划的可行域为凸集，其局部最优解即为全局最优解，而且其最优解的集合形成一个凸集。当凸规划的目标函数 $f(\boldsymbol{X})$ 为严格凸函数时，其最优解必定唯一(假定最优解存在)。由此可见，凸规划是一类比较简单而又具有重要理论意义的非线性规划。

由于线性函数既可视为凸函数，又可视为凹函数，故线性规划也属于凸规划。

例 4 试分析非线性规划

$$\begin{cases} \min\ f(\boldsymbol{X}) = x_1^2 + x_2^2 - 4x_1 + 4 \\ g_1(\boldsymbol{X}) = x_1 - x_2 + 2 \geqslant 0 \\ g_2(\boldsymbol{X}) = -x_1^2 + x_2 - 1 \geqslant 0 \\ x_1, x_2 \geqslant 0 \end{cases}$$

解 $f(\boldsymbol{X})$ 和 $g_2(\boldsymbol{X})$ 的海赛矩阵的行列式分别是

$$|\boldsymbol{H}| = \begin{vmatrix} \dfrac{\partial^2 f(\boldsymbol{X})}{\partial x_1^2} & \dfrac{\partial^2 f(\boldsymbol{X})}{\partial x_1 \partial x_2} \\ \dfrac{\partial^2 f(\boldsymbol{X})}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f(\boldsymbol{X})}{\partial x_2^2} \end{vmatrix}$$

$$= \begin{vmatrix} 2 & 0 \\ 0 & 2 \end{vmatrix} = 4 > 0$$

$$|g_2|=\begin{vmatrix}\dfrac{\partial^2 g_2(\boldsymbol{X})}{\partial x_1^2} & \dfrac{\partial^2 g_2(\boldsymbol{X})}{\partial x_1\partial x_2}\\ \dfrac{\partial^2 g_2(\boldsymbol{X})}{\partial x_2\partial x_1} & \dfrac{\partial^2 g_2(\boldsymbol{X})}{\partial x_2^2}\end{vmatrix}$$

$$=\begin{vmatrix}-2 & 0\\ 0 & 0\end{vmatrix}=0$$

知 $f(\boldsymbol{X})$为严格凸函数，$g_2(\boldsymbol{X})$为凹函数。由于其他约束条件均为线性函数，所以这是一个凸规划问题(见图 7-4)。C 点为其最优点：$\boldsymbol{X}^*=(0.58,1.34)^{\mathrm{T}}$，目标函数的最优值为$f(\boldsymbol{X}^*)=3.8$。

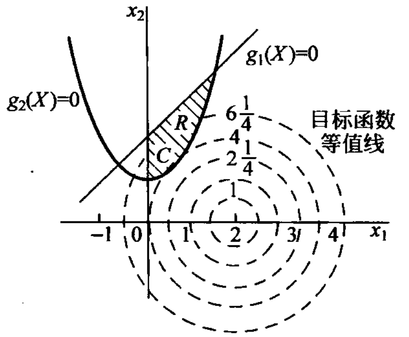


图 7-4

7.1.5 下降迭代算法

为了求某可微函数(假定无约束)的最优解，根据前面的叙述，可如下进行：令该函数的梯度等于零，由此求得平稳点；然后用充分条件进行判别，求出所要的解。对某些较简单的函数，这样做有时是可行的；但对一般 n 元函数 $f(\boldsymbol{X})$来说，由条件$\nabla f(\boldsymbol{X})=0$ 得到的常是一个非线性方程组，解它相当困难。对于不可微函数，当然谈不上使用这样的方法。为此，常直接使用**迭代法**。

迭代法的基本思想是：为了求函数 $f(\boldsymbol{X})$的最优解，首先给定一个**初始估计** $\boldsymbol{X}^{(0)}$，然后按某种规则(即**算法**)找出比 $\boldsymbol{X}^{(0)}$ 更好的解 $\boldsymbol{X}^{(1)}$(对极小化问题，$f(\boldsymbol{X}^{(1)})<f(\boldsymbol{X}^{(0)})$；对极大化问题，$f(\boldsymbol{X}^{(1)})>f(\boldsymbol{X}^{(0)})$)，再按此种规则找出比 $\boldsymbol{X}^{(1)}$ 更好的解 $\boldsymbol{X}^{(2)}$……。如此即可得到一个解的序列$\{\boldsymbol{X}^{(k)}\}$。若这个解序列有极限 $\boldsymbol{X}^*$，即

$$\lim_{k\to\infty}\|\boldsymbol{X}^{(k)}-\boldsymbol{X}^*\|=0,$$

则称它**收敛**于 $\boldsymbol{X}^*$。

若这算法是有效的，那么它所产生的解的序列将收敛于该问题的最优解。不过，由于计算机只能进行有限次迭代，一般说很难得到准确解，而只能得到近似解。当满足所要求的精度时，即可停止迭代。

若由某算法所产生的解的序列$\{\boldsymbol{X}^{(k)}\}$使目标函数值 $f(\boldsymbol{X}^{(k)})$逐步减少，就称这算法为**下降算法**。"下降"的要求比较容易实现，它包含了很多种具体算法。显然，求解极小化问题应采用下降算法。

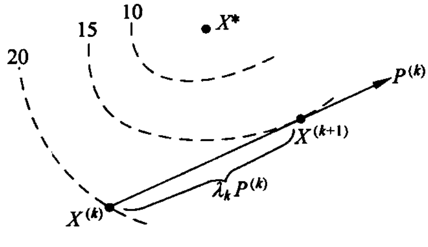


图 7-5

现假定已迭代到点 $\boldsymbol{X}^{(k)}$(见图 7-5)，若从 $\boldsymbol{X}^{(k)}$ 出发沿任何方向移动都不能使目标函数值下降，则 $\boldsymbol{X}^{(k)}$ 是一局部极小点，迭代停止。若从 $\boldsymbol{X}^{(k)}$ 出发至少存在一个方向可使目标函数值有所下降，则可选定能使目标函数值下降的某方向 $\boldsymbol{P}^{(k)}$，沿这个方向迈进适当的一步，得到下一个迭代点 $\boldsymbol{X}^{(k+1)}$，

并使 $f(\boldsymbol{X}^{(k+1)})<f(\boldsymbol{X}^{(k)})$。这相当于在射线

$$\boldsymbol{X}=\boldsymbol{X}^{(k)}+\lambda\boldsymbol{P}^{(k)}$$

上选定新点

$$\boldsymbol{X}^{(k+1)}=\boldsymbol{X}^{(k)}+\lambda_k\boldsymbol{P}^{(k)}$$

其中，$\boldsymbol{P}^{(k)}$ 称为**搜索方向**；λ_k 称为**步长**或步长因子。

下降迭代算法的步骤可总结如下：

(1) 选定某一初始点 $\boldsymbol{X}^{(0)}$，并令 $k:=0$；

(2) 确定搜索方向 $\boldsymbol{P}^{(k)}$；

(3) 从 $\boldsymbol{X}^{(k)}$ 出发，沿方向 $\boldsymbol{P}^{(k)}$ 求步长 λ_k，以产生下一个迭代点 $\boldsymbol{X}^{(k+1)}$；

(4) 检查得到的新点 $\boldsymbol{X}^{(k+1)}$ 是否为极小点或近似极小点。若是，则停止迭代。否则，令 $k:=k+1$，转回(2)继续进行迭代。

在以上步骤中，选取搜索方向 $\boldsymbol{P}^{(k)}$ 是最关键的一步，有关各种算法的区分，主要在于确定搜索方向的方法不同。

确定步长 λ_k 可选用不同的方法。最简单的一种是令它等于某一常数(例如令 $\lambda_k=1$)，这样做计算简便，但不能保证使目标函数值下降。第二种称为可接受点算法，只要能使目标函数值下降，可任意选取步长 λ_k。第三种方法是基于沿搜索方向使目标函数值下降最多，即沿射线 $\boldsymbol{X}=\boldsymbol{X}^{(k)}+\lambda\boldsymbol{P}^{(k)}$ 求目标函数 $f(\boldsymbol{X})$ 的极小(注意，这里是指无约束问题)

$$\lambda_k:\ \min f(\boldsymbol{X}^{(k)}+\lambda\boldsymbol{P}^{(k)}) \tag{7-20}$$

由于这项工作是求以 λ 为变量的一元函数 $f(\boldsymbol{X}^{(k)}+\lambda\boldsymbol{P}^{(k)})$ 的极小点 λ_k，故常称这一过程为(最优)**一维搜索**或线搜索，这样确定的步长为**最佳步长**。

一维搜索有个十分重要的性质：在搜索方向上所得最优点处目标函数的梯度和该搜索方向正交。

定理 7 设目标函数 $f(\boldsymbol{X})$ 具有一阶连续偏导数，$\boldsymbol{X}^{(k+1)}$ 按下述规则产生

$$\begin{cases}\lambda_k:\ \min\limits_{\lambda} f(\boldsymbol{X}^{(k)}+\lambda\boldsymbol{P}^{(k)})\\ \boldsymbol{X}^{(k+1)}=\boldsymbol{X}^{(k)}+\lambda_k\boldsymbol{P}^{(k)}\end{cases}$$

则有

$$\nabla f(\boldsymbol{X}^{(k+1)})^{\mathrm{T}}\boldsymbol{P}^{(k)}=0 \tag{7-21}$$

证明 构造函数 $\varphi(\lambda)=f(\boldsymbol{X}^{(k)}+\lambda\boldsymbol{P}^{(k)})$，则得

$$\begin{cases}\varphi(\lambda_k)=\min\limits_{\lambda}\varphi(\lambda)\\ \boldsymbol{X}^{(k+1)}=\boldsymbol{X}^{(k)}+\lambda_k\boldsymbol{P}^{(k)}\end{cases}$$

即 λ_k 为 $\varphi(\lambda)$ 的极小点。此外 $\varphi'(\lambda)=\nabla f(\boldsymbol{X}^{(k)}+\lambda\boldsymbol{P}^{(k)})^{\mathrm{T}}\boldsymbol{P}^{(k)}$。

由 $\varphi'(\lambda)|_{\lambda=\lambda_k}=0$，可得

$$\nabla f(\boldsymbol{X}^{(k)}+\lambda_k\boldsymbol{P}^{(k)})^{\mathrm{T}}\boldsymbol{P}^{(k)}=\nabla f(\boldsymbol{X}^{(k+1)})^{\mathrm{T}}\boldsymbol{P}^{(k)}=0$$

定理得证。

式(7-21)的几何意义见图 7-6。

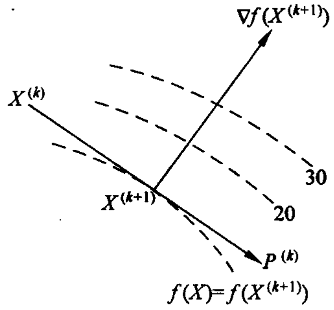


图 7-6

对一个好的算法,不仅要求它产生的点列能收敛到问题的最优解,还要求具有较快的**收敛速度**。设序列$\{X^{(k)}\}$收敛于X^*,若存在与迭代次数k无关的数$0<\beta<\infty$和$\alpha\geqslant 1$,使k从某个$k_0>0$开始都有

$$\| X^{(k+1)}-X^* \| \leqslant \beta \| X^{(k)}-X^* \|^{\alpha} \tag{7-22}$$

成立,就称$\{X^{(k)}\}$收敛的阶为α,或$\{X^{(k)}\}$$\alpha$阶收敛。

当$\alpha=2$时,称为**二阶收敛**,也可说$\{X^{(k)}\}$具有二阶敛速。

当$1<\alpha<2$时,**称超线性收敛**。

当$\alpha=1$,且$0<\beta<1$时,**称线性收敛**或一阶收敛。

一般来讲,线性收敛的敛速是比较慢的,二阶收敛是很快的,超级性收敛介于以上两者之间。若一个算法具有超线性或更高的收敛速度,就认为它是一个很好的算法。

因为真正的最优解事先并不知道,为决定什么时候停止计算,只能根据相继两次迭代的结果。常用的终止计算准则有以下几种。

(1) 根据相继两次迭代的**绝对误差**

$$\| X^{(k+1)}-X^{(k)} \| < \varepsilon_1 \tag{7-23}$$

$$| f(X^{(k+1)})-f(X^{(k)}) | < \varepsilon_2 \tag{7-24}$$

(2) 根据相继两次迭代的**相对误差**

$$\frac{\| X^{(k+1)}-X^{(k)} \|}{\| X^{(k)} \|} < \varepsilon_3 \tag{7-25}$$

$$\frac{| f(X^{(k+1)})-f(X^{(k)}) |}{| f(X^{(k)}) |} < \varepsilon_4 \tag{7-26}$$

这时要求分母不等于和不接近于零。

(3) 根据目标函数梯度的模足够小

$$\| \nabla f(X^{(k)}) \| < \varepsilon_5 \tag{7-27}$$

其中,$\varepsilon_1,\varepsilon_2,\varepsilon_3,\varepsilon_4,\varepsilon_5$为事先给定的足够小的正数。

7.2 一维搜索

前已述及,当用上述迭代法求函数的极小点时,常常要用到一维搜索,即沿某一已知方向求目标函数的极小点。一维搜索的方法很多,常用的有:

(1) 试探法(“成功一失败”法,斐波那契法,0.618 法等);

(2) 插值法(抛物线插值法,三次插值法等);

(3) 微积分中的求根法(切线法,二分法等)。

限于篇幅，以下仅介绍斐波那契法和 0.618 法。

7.2.1 斐波那契(Fibonacci)法(分数法)

设 $y=f(t)$ 是区间 $[a,b]$ 上的**下单峰函数**(图 7-7)，在此区间内它有唯一极小点 t^*。若在此区间内任取两点 a_1 和 b_1，$a_1<b_1$，并计算函数值 $f(a_1)$ 和 $f(b_1)$，可能出现以下两种情形：

(1) $f(a_1)<f(b_1)$(图 7-7(a))，这时极小点 t^* 必在区间 $[a,b_1]$ 内。

(2) $f(a_1)\geqslant f(b_1)$(图 7-7(b))，这时极小点 t^* 必在区间 $[a_1,b]$ 内。

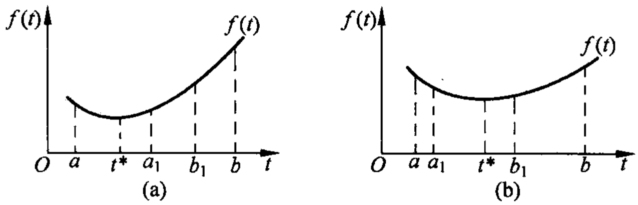


图 7-7

这说明，只要在区间 $[a,b]$ 内取两个不同点，并算出它们的函数值加以比较，就可以把**搜索区间** $[a,b]$ 缩小成 $[a,b_1]$ 或 $[a_1,b]$(缩小后的区间仍需包含极小点)。现在，如果要继续缩小搜索区间 $[a,b_1]$(或 $[a_1,b]$)，就只需在上述区间内再取一点算出其函数值，并与 $f(a_1)$ 或 $f(b_1)$ 加以比较即可。只要缩小后的区间包含极小点 t^*，则区间缩小得越小，就越接近于函数的极小点，但计算函数值的次数也就越多。这就说明区间的**缩短率**和函数值的计算次数有关。现在要问，计算函数值 n 次，能把包含有极小点的区间缩小到什么程度呢？或者换一种说法，计算函数值 n 次能把原来多大的区间缩小成长度为一个单位的区间呢？

如果用 F_n 表示计算 n 个函数值能缩短为单位区间的最大原区间长度，显然

$$F_0=F_1=1 \tag{7-28}$$

其原因是，只有当原区间长度本来就是一个单位长度时才不必计算函数值；此外，只计算一次函数值无法将区间缩短，故只有区间长度本来就是单位区间时才行。

现考虑计算函数值两次的情形，今后我们把计算函数值的点称做**试算点**或试点。

在区间 $[a,b]$ 内取两个不同点 a_1 和 b_1(图 7-8(a))，计算其函数值以缩短区间，缩短后的区间为 $[a,b_1]$ 或 $[a_1,b]$。显然，这两个区间长度之和必大于 $[a,b]$ 的长度，也就是说，计算两次函数值一般无法把长度大于两个单位的区间缩成单位区间。但是，对于长度为两个单位的区间，可以如图 7-8(b)那样选取试点 a_1 和 b_1，图中 ε 为任意小的正数，缩短后的区间长度为 $1+\varepsilon$。由于 ε 可任意选取，故缩短后的区间长度接近于一个单位长度。由此可得 $F_2=2$。

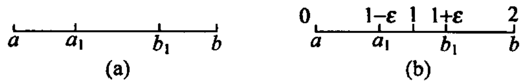


图 7-8

根据同样的分析(见图 7-9)可得

$$F_3=3,\quad F_4=5,\quad F_5=8,\cdots$$

序列$\{F_n\}$可写成一个递推公式

$$F_n=F_{n-1}+F_{n-2}\qquad n\geqslant 2 \tag{7-29}$$

利用式(7-29),可依次算出各 F_n 的值,见表 7-1。这些 F_n 就是通常所说的斐波那契数。

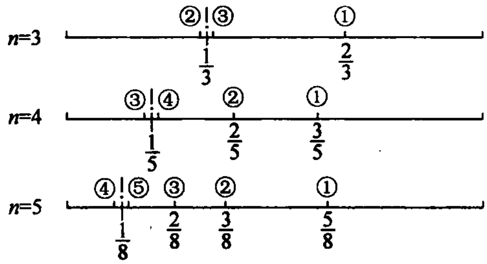


图 7-9

表 7-1

n	0	1	2	3	4	5	6	7	8	9	10	11	12
F_n	1	1	2	3	5	8	13	21	34	55	89	144	233

由以上讨论可知,计算 n 次函数值所能获得的最大缩短率(缩短后的区间长度与原区间长度之比)为 $1/F_n$。例如 $F_{20}=10\,946$,所以计算 20 个函数值即可把原长度为 L 的区间缩短为

$$\frac{L}{10\,946}=0.000\,09L$$

的区间。现在,要想计算 n 个函数值,而把区间$[a_0,b_0]$的长度缩短为原来长度的 δ 倍,即缩短后的区间长度为

$$b_{n-1}-a_{n-1}\leqslant(b_0-a_0)\delta$$

则只要 n 足够大,能使下式成立即可

$$F_n\geqslant\frac{1}{\delta} \tag{7-30}$$

其中,δ 为一个正小数,称为区间缩短的相对精度。有时给出区间缩短的绝对精度 η,即

要求

$$b_{n-1}-a_{n-1}\leqslant\eta \tag{7-31}$$

显然，上述相对精度和绝对精度之间有如下关系

$$\eta=(b_0-a_0)\delta \tag{7-32}$$

用这个方法缩短区间的步骤如下：

(1) 确定试点的个数 n。根据相对精度 δ，即可用式(7-30)算出 F_n，然后由表 7-1 确定最小的 n。

(2) 选取前两个试点的位置。

由式(7-29)可知第一次缩短时的两个试点位置分别是(见图 7-10)：

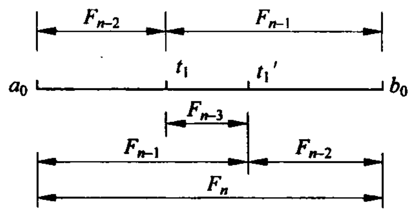


图　7-10

$$\begin{cases} t_1=a_0+\dfrac{F_{n-2}}{F_n}(b_0-a_0) \\ \quad =b_0+\dfrac{F_{n-1}}{F_n}(a_0-b_0) \\ t'_1=a_0+\dfrac{F_{n-1}}{F_n}(b_0-a_0) \end{cases} \tag{7-33}$$

它们在区间内的位置是对称的。

(3) 计算函数值 $f(t_1)$和 $f(t'_1)$，并比较它们的大小。

若 $f(t_1)<f(t'_1)$，则取

$$a_1=a_0 \qquad b_1=t'_1 \qquad t'_2=t_1$$

并令

$$t_2=b_1+\frac{F_{n-2}}{F_{n-1}}(a_1-b_1)$$

否则，取

$$a_1=t_1 \qquad b_1=b_0 \qquad t_2=t'_1$$

并令

$$t'_2=a_1+\frac{F_{n-2}}{F_{n-1}}(b_1-a_1)$$

(4) 计算 $f(t_2)$或 $f(t'_2)$(其中的一个已经算出)，如第 3 步那样一步步迭代。计算试点的一般公式为

$$\begin{cases} t_k=b_{k-1}+\dfrac{F_{n-k}}{F_{n-k+1}}(a_{k-1}-b_{k-1}) \\ t'_k=a_{k-1}+\dfrac{F_{n-k}}{F_{n-k+1}}(b_{k-1}-a_{k-1}) \end{cases} \tag{7-34}$$

其中，$k=1,2,\cdots,n-1$。

(5) 当进行至 $k=n-1$ 时

$$t_{n-1}=t'_{n-1}=\frac{1}{2}(a_{n-2}+b_{n-2})$$

这就无法借比较函数值 $f(t_{n-1})$ 和 $f(t'_{n-1})$ 的大小以确定最终区间，为此，取

$$\begin{cases} t_{n-1}=\dfrac{1}{2}(a_{n-2}+b_{n-2}) \\ t'_{n-1}=a_{n-2}+\left(\dfrac{1}{2}+\varepsilon\right)(b_{n-2}-a_{n-2}) \end{cases} \tag{7-35}$$

其中 ε 为任意小的数。在 t_{n-1} 和 t'_{n-1} 这两点中，以函数值较小者为近似极小点，相应的函数值为近似极小值，并得最终区间 $[a_{n-2},t'_{n-1}]$ 或 $[t_{n-1},b_{n-2}]$。

由上述分析可知，斐波那契法使用**对称搜索**的方法，逐步缩短所考察的区间，它能以尽量少的函数求值次数，达到预定的某一缩短率。

例 5 试用斐波那契法求函数 $f(t)=t^2-t+2$ 的近似极小点和极小值，要求缩短后的区间长度不大于区间 $[-1,3]$ 的 0.08 倍。

解 容易验证，在此区间上函数 $f(t)=t^2-t+2$ 为严格凸函数。为了进行比较，我们给出其精确解是：$t^*=0.5$，$f(t^*)=1.75$。

已知 $\delta=0.08$，$F_n\geqslant 1/\delta=1/0.08=12.5$

查表 7-1，$n=6$，$a_0=-1$，$b_0=3$

$$t_1=b_0+\frac{F_5}{F_6}(a_0-b_0)=3+\frac{8}{13}(-1-3)=0.538$$

$$t'_1=a_0+\frac{F_5}{F_6}(b_0-a_0)=-1+\frac{8}{13}(3-(-1))=1.462$$

$$f(t_1)=0.538^2-0.538+2=1.751$$

$$f(t'_1)=1.462^2-1.462+2=2.675$$

由于 $f(t_1)<f(t'_1)$，故取 $a_1=-1$，$b_1=1.462$，$t'_2=0.538$

$$t_2=b_1+\frac{F_4}{F_5}(a_1-b_1)=1.462+\frac{5}{8}(-1-1.462)=-0.077$$

$$f(t_2)=(-0.077)^2-(-0.077)+2=2.083$$

由于 $f(t_2)>f(t'_2)=1.751$，故取 $a_2=-0.077$，$b_2=1.462$，$t_3=0.538$

$$t'_3=a_2+\frac{F_3}{F_4}(b_2-a_2)=-0.077+\frac{3}{5}(1.462+0.077)=0.846$$

$$f(t'_3)=0.846^2-0.846+2=1.870$$

由于 $f(t'_3)>f(t_3)=1.751$，故取 $a_3=-0.077$，$b_3=0.846$，$t'_4=0.538$

$$t_4=b_3+\frac{F_2}{F_3}(a_3-b_3)=0.846+\frac{2}{3}(-0.077-0.846)=0.231$$

$$f(t_4)=0.231^2-0.231+2=1.822$$

由于 $f(t_4)>f(t'_4)=1.751$，故取 $a_4=0.231$，$b_4=0.846$，$t_5=0.538$。现令 $\varepsilon=0.01$，则

$$t'_5 = a_4 + \left(\frac{1}{2} + \varepsilon\right)(b_4 - a_4)$$

$$= 0.231 + (0.5 + 0.01)(0.846 - 0.231) = 0.545$$

$$f(t'_5) = 0.545^2 - 0.545 + 2 = 1.752 > f(t_5) = 1.751$$

故取 $a_5 = 0.231, b_5 = 0.545$。由于 $f(t_5) = 1.751 < f(t'_5) = 1.752$，所以以 t_5 为近似极小点，近似极小值为 1.751。

缩短后的区间长度为 $0.545 - 0.231 = 0.314$， $0.314/4 = 0.0785 < 0.08$，其整个计算过程示于图 7-11 中。

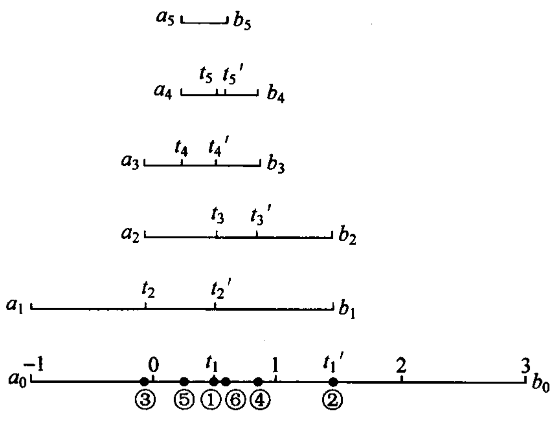


图 7-11

7.2.2 0.618 法(黄金分割法)

由上节的论述可知，当用斐波那契法以 n 个试点来缩短某一区间时，区间长度的第一次缩短率为 F_{n-1}/F_n，其后各次分别为

$$\frac{F_{n-2}}{F_{n-1}},\quad \frac{F_{n-3}}{F_{n-2}},\quad \cdots,\quad \frac{F_1}{F_2}$$

现将以上数列分为奇数项 F_{2k-1}/F_{2k} 和偶数项 F_{2k}/F_{2k+1}，可以证明，这两个数列收敛于同一个极限。

设当 $k\to\infty$ 时

$$\frac{F_{2k-1}}{F_{2k}} \to \lambda \qquad \frac{F_{2k}}{F_{2k+1}} \to \mu$$

由于

$$\frac{F_{2k-1}}{F_{2k}} = \frac{F_{2k-1}}{F_{2k-1} + F_{2k-2}} = \frac{1}{1 + \dfrac{F_{2k-2}}{F_{2k-1}}}$$

故当 $k\to\infty$ 时

$$\lim_{k\to\infty}\frac{F_{2k-1}}{F_{2k}}=\frac{1}{1+\mu}=\lambda \tag{7-36}$$

同理可证

$$\mu=\frac{1}{1+\lambda} \tag{7-37}$$

将式(7-36)代入式(7-37)得

$$\mu=\frac{1+\mu}{2+\mu}$$

即

$$\mu^2+\mu-1=0$$

从而可得

$$\mu=\frac{\sqrt{5}-1}{2}$$

若把式(7-37)代入式(7-36),则得

$$\lambda^2+\lambda-1=0$$

故有

$$\lambda=\mu=\frac{\sqrt{5}-1}{2}=0.618\,033\,988\,741\,894\,8 \tag{7-38}$$

现用不变的区间缩短率0.618,代替斐波那契法每次不同的缩短率,就得到了黄金分割法(0.618法)。这个方法可以看成是斐波那契法的近似,实现起来比较容易,效果也相当好,因而易于为人们所接受。

当用0.618方法时,计算 n 个试点的函数值可以把原区间$[a_0,b_0]$连续缩短$n-1$次,因为每次的缩短率均为 μ,故最后的区间长度为

$$(b_0-a_0)\mu^{n-1}$$

这就是说,当已知缩短的相对精度为 δ 时,可用下式计算试点个数 n

$$\mu^{n-1}\leqslant\delta \tag{7-39}$$

当然,也可以不预先计算试点的数目 n,而在计算过程中逐次加以判断,看是否已满足了提出的精度要求。

0.618法是一种**等速对称**进行试探的方法,每次的试点均取在区间长度的0.618倍和0.382倍处。

7.3 无约束极值问题的解法

本节研究无约束极值问题的解法,这种问题可表述为

$$\min f(\boldsymbol{X}),\quad \boldsymbol{X}\in E^n \tag{7-40}$$

前面曾指出,在求解上述问题时常使用迭代法,迭代法可大体分为两大类。一类要用到函

数的一阶导数和(或)二阶导数,由于用到了函数的解析性质,故称为**解析法**;另一类在迭代过程中仅用到函数值,而不要求函数的解析性质,这类方法称为**直接法**。一般来说,直接法的收敛速度较慢,只是在变量较少时才适用。但是直接法的迭代步骤简单,特别是当目标函数的解析表达式十分复杂,甚至写不出具体表达式时,它们的导数很难求得,或者根本不存在,这时解析法就无能为力了。

本节仅介绍几种常用的基本方法,其中前三种属解析法,后面一种属直接法。

7.3.1 梯度法(最速下降法)

在求解无约束极值问题的解析法中,**梯度法**是最为古老但又十分基本的一种数值方法。它的迭代过程简单,使用方便,而且又是理解某些其他最优化方法的基础,所以我们先来说明这一方法。

1. 梯度法的基本原理

假定无约束极值问题式(7-40)中的目标函数 $f(\boldsymbol{X})$ 有一阶连续偏导数,具有极小点 $\boldsymbol{X}^*$。以 $\boldsymbol{X}^{(k)}$ 表示极小点的第 k 次近似,为了求其第 $k+1$ 次近似点 $\boldsymbol{X}^{(k+1)}$,我们在 $\boldsymbol{X}^{(k)}$ 点沿方向 $\boldsymbol{P}^{(k)}$ 作射线

$$\boldsymbol{X}=\boldsymbol{X}^{(k)}+\lambda\boldsymbol{P}^{(k)}\quad(\lambda\geqslant 0)$$

现将 $f(\boldsymbol{X})$ 在 $\boldsymbol{X}^{(k)}$ 点处展成泰勒级数

$$\begin{aligned}f(\boldsymbol{X})&=f(\boldsymbol{X}^{(k)}+\lambda\boldsymbol{P}^{(k)})\\&=f(\boldsymbol{X}^{(k)})+\lambda\nabla f(\boldsymbol{X}^{(k)})^{\mathrm{T}}\boldsymbol{P}^{(k)}+o(\lambda)\end{aligned}$$

其中

$$\lim_{\lambda\to 0}\frac{o(\lambda)}{\lambda}=0$$

对于充分小的 λ,只要

$$\nabla f(\boldsymbol{X}^{(k)})^{\mathrm{T}}\boldsymbol{P}^{(k)}<0\qquad(7\text{-}41)$$

即可保证 $f(\boldsymbol{X}^{(k)}+\lambda\boldsymbol{P}^{(k)})<f(\boldsymbol{X}^{(k)})$。这时若取

$$\boldsymbol{X}^{(k+1)}=\boldsymbol{X}^{(k)}+\lambda\boldsymbol{P}^{(k)}$$

就能使目标函数值得到改善。

现考查不同的方向 $\boldsymbol{P}^{(k)}$。假定 $\boldsymbol{P}^{(k)}$ 的模一定(且不为零),并设 $\nabla f(\boldsymbol{X}^{(k)})\neq 0$(否则,$\boldsymbol{X}^{(k)}$ 是平稳点),使式(7-41)成立的 $\boldsymbol{P}^{(k)}$ 有无限多个。为了使目标函数值能得到尽量大的改善,必须寻求使 $\nabla f(\boldsymbol{X}^{(k)})^{\mathrm{T}}\boldsymbol{P}^{(k)}$ 取最小值的 $\boldsymbol{P}^{(k)}$。由线性代数学知道

$$\nabla f(\boldsymbol{X}^{(k)})^{\mathrm{T}}\boldsymbol{P}^{(k)}=\|\nabla f(\boldsymbol{X}^{(k)})\|\cdot\|\boldsymbol{P}^{(k)}\|\cos\theta\qquad(7\text{-}42)$$

式中 θ 为向量 $\nabla f(\boldsymbol{X}^{(k)})$ 与 $\boldsymbol{P}^{(k)}$ 的夹角。当 $\boldsymbol{P}^{(k)}$ 与 $\nabla f(\boldsymbol{X}^{(k)})$ 反向时,$\theta=180°$,$\cos\theta=-1$。这时式(7-41)成立,而且其左端取最小值。我们称方向

$$\boldsymbol{P}^{(k)}=-\nabla f(\boldsymbol{X}^{(k)})$$

为**负梯度方向**，它是使函数值下降最快的方向(在 $X^{(k)}$ 的某一小范围内)。

为了得到下一个近似极小点，在选定了搜索方向之后，还要确定步长 λ。当采用可接受点算法时，就是取某一 λ 进行试算，看是否满足不等式

$$f(X^{(k)}-\lambda\nabla f(X^{(k)}))<f(X^{(k)}) \tag{7-43}$$

若上述不等式成立，就可以迭代下去。否则，缩小 λ 使满足不等式式(7-43)。由于采用负梯度方向，满足式(7-43)的 λ 总是存在的。

另一种方法是通过在负梯度方向的一维搜索，来确定使 $f(X)$ 最小的 λ_k，这种梯度法就是所谓**最速下降法**。

2. 计算步骤

现将用梯度法解无约束极值问题的步骤简要总结如下：

(1) 给定初始近似点 $X^{(0)}$ 及精度 $\varepsilon>0$，若 $\|\nabla f(X^{(0)})\|^2\leqslant\varepsilon$，则 $X^{(0)}$ 即为近似极小点。

(2) 若 $\|\nabla f(X^{(0)})\|^2>\varepsilon$，求步长 λ_0，并计算

$$X^{(1)}=X^{(0)}-\lambda_0\nabla f(X^{(0)})$$

求步长可用一维搜索法、微分法或试算法。若求最佳步长，则应使用前两种方法。

(3) 一般地，设已迭代到点 $X^{(k)}$，若 $\|\nabla f(X^{(k)})\|^2\leqslant\varepsilon$，则 $X^{(k)}$ 即为所求的近似解；若 $\|\nabla f(X^{(k)})\|^2>\varepsilon$，则求步长 λ_k，并确定下一个近似点

$$X^{(k+1)}=X^{(k)}-\lambda_k\nabla f(X^{(k)}) \tag{7-44}$$

如此继续，直至达到要求的精度为止。

若 $f(X)$ 具有二阶连续偏导数，在 $X^{(k)}$ 作 $f(X^{(k)}-\lambda\nabla f(X^{(k)}))$ 的泰勒展开

$$f(X^{(k)}-\lambda\nabla f(X^{(k)}))\approx f(X^{(k)})-\nabla f(X^{(k)})^{\mathrm{T}}\lambda\nabla f(X^{(k)})+\frac{1}{2}\lambda\nabla f(X^{(k)})^{\mathrm{T}}H(X^{(k)})\lambda\nabla f(X^{(k)})$$

对 λ 求导并令其等于零，则得**近似最佳步长**

$$\lambda_k=\frac{\nabla f(X^{(k)})^{\mathrm{T}}\nabla f(X^{(k)})}{\nabla f(X^{(k)})^{\mathrm{T}}H(X^{(k)})\nabla f(X^{(k)})} \tag{7-45}$$

可见近似最佳步长不只与梯度有关，而且与海赛矩阵 H 也有关系，计算起来比较麻烦。

确定步长 λ_k 也可不用式(7-45)，而采用任一种一维搜索法(例如 0.618 法等)。

有时，将搜索方向 $P^{(k)}$ 的模规格化为 1，在这种情况下

$$P^{(k)}=\frac{-\nabla f(X^{(k)})}{\|\nabla f(X^{(k)})\|} \tag{7-46}$$

同时，式(7-45)变为

$$\lambda_k=\frac{\nabla f(X^{(k)})^{\mathrm{T}}\nabla f(X^{(k)})\|\nabla f(X^{(k)})\|}{\nabla f(X^{(k)})^{\mathrm{T}}H(X^{(k)})\nabla f(X^{(k)})} \tag{7-47}$$

例 6 试用梯度法求

$$f(\boldsymbol{X})=(x_1-1)^2+(x_2-1)^2$$

的极小点，已知 $\varepsilon=0.1$。

解 取初始点 $\boldsymbol{X}^{(0)}=(0,0)^{\mathrm{T}}$

$$\nabla f(\boldsymbol{X})=[2(x_1-1),2(x_2-1)]^{\mathrm{T}}$$

$$\nabla f(\boldsymbol{X}^{(0)})=(-2,-2)^{\mathrm{T}}$$

$$\|\nabla f(\boldsymbol{X}^{(0)})\|^2=(\sqrt{(-2)^2+(-2)^2})^2=8>\varepsilon$$

$$\boldsymbol{H}(\boldsymbol{X})=\begin{pmatrix}2&0\\0&2\end{pmatrix}$$

由式(7-45)

$$\lambda_0=\frac{\nabla f(\boldsymbol{X}^{(0)})^{\mathrm{T}}\nabla f(\boldsymbol{X}^{(0)})}{\nabla f(\boldsymbol{X}^{(0)})^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{X}^{(0)})\nabla f(\boldsymbol{X}^{(0)})}$$

$$=\frac{(-2,-2)\begin{pmatrix}-2\\-2\end{pmatrix}}{(-2,-2)\begin{pmatrix}2&0\\0&2\end{pmatrix}\begin{pmatrix}-2\\-2\end{pmatrix}}=\frac{8}{16}=\frac{1}{2}$$

$$\boldsymbol{X}^{(1)}=\boldsymbol{X}^{(0)}-\lambda_0\nabla f(\boldsymbol{X}^{(0)})=\begin{pmatrix}0\\0\end{pmatrix}-\frac{1}{2}\begin{pmatrix}-2\\-2\end{pmatrix}=\begin{pmatrix}1\\1\end{pmatrix}$$

$$\nabla f(\boldsymbol{X}^{(1)})=[2(1-1),2(1-1)]^{\mathrm{T}}=(0,0)^{\mathrm{T}}$$

故 $\boldsymbol{X}^{(1)}$ 即为极小点。

注意，计算步长 λ_0 时也可不用海赛矩阵。由于

$$\boldsymbol{X}^{(1)}=\boldsymbol{X}^{(0)}-\lambda\nabla f(\boldsymbol{X}^{(0)})=\begin{pmatrix}0\\0\end{pmatrix}-\lambda\begin{pmatrix}-2\\-2\end{pmatrix}=\begin{pmatrix}2\lambda\\2\lambda\end{pmatrix}$$

代入目标函数可得

$$f(\boldsymbol{X}^{(1)})=(2\lambda-1)^2+(2\lambda-1)^2=2(2\lambda-1)^2$$

令

$$\mathrm{d}f(\boldsymbol{X}^{(1)})/\mathrm{d}\lambda=0$$

即得所求步长 $\lambda_0=1/2$。

由这个例子可知，对于目标函数的等值线为圆的问题来说，不管初始点位置取在哪里，负梯度方向总是直指圆心，而圆心即为极值点。这样，只要一次迭代即可达到最优解。

例 7 试求 $f(\boldsymbol{X})=x_1^2+25x_2^2$ 的极小点。

解 取初始点 $\boldsymbol{X}^{(0)}=(2,2)^{\mathrm{T}}$，$f(\boldsymbol{X}^{(0)})=104$。本例使用**规格化搜索方向法**。

现先取用固定步长 $\lambda=1$，其迭代过程示如表 7-2。

表 7-2

步骤	点	x_1	x_2	$\frac{\partial f(\boldsymbol{X}^{(k)})}{\partial x_1}$	$\frac{\partial f(\boldsymbol{X}^{(k)})}{\partial x_2}$	$\Vert \nabla f(\boldsymbol{X}^{(k)}) \Vert$
0	$\boldsymbol{X}^{(0)}$	2	2	4	100	~100
1	$\boldsymbol{X}^{(1)}$	1.96	1.00	3.92	50	50.1
2	$\boldsymbol{X}^{(2)}$	1.88	0	3.76	0	3.76
3	$\boldsymbol{X}^{(3)}$	0.88	0	1.76	0	1.76
4	$\boldsymbol{X}^{(4)}$	−0.12	0	−0.24	0	0.24
5	$\boldsymbol{X}^{(5)}$	0.88	0			

继续计算下去可以看出，x_1 将来回振荡，难以收敛到极小点(0,0)。为使迭代过程收敛，必须不断减小步长 λ 的值。

采用最佳步长时收敛较快，而且相邻两步的搜索方向互相垂直。下面用最佳步长进行搜索。其迭代过程列于表 7-3 中。

表 7-3

步骤	点	λ_k	x_1	x_2	$\frac{\partial f(\boldsymbol{X}^{(k)})}{\partial x_1}$	$\frac{\partial f(\boldsymbol{X}^{(k)})}{\partial x_2}$	$f(\boldsymbol{X}^{(k)})$
0	$\boldsymbol{X}^{(0)}$	2.003	2	2	4	100	104
1	$\boldsymbol{X}^{(1)}$	1.850	1.92	−0.003	3.84	−0.15	3.69
2	$\boldsymbol{X}^{(2)}$	0.070	0.070	0.070	0.14	3.50	0.13
3	$\boldsymbol{X}^{(3)}$		0.070	−0.000			

为直观起见，将上述两种迭代过程分别画于图 7-12 及图 7-13 中。

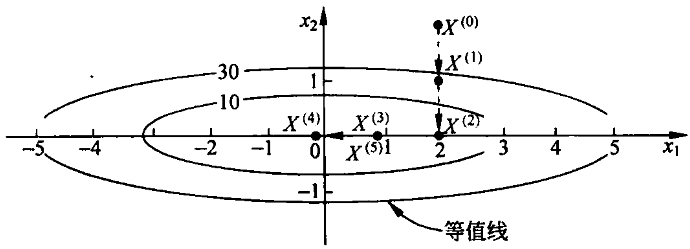


图 7-12

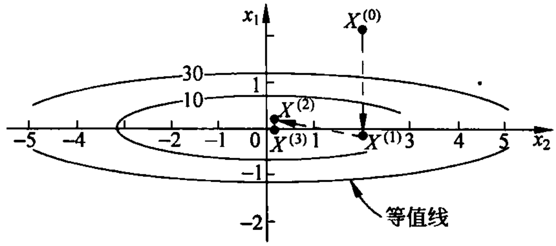


图 7-13

可以证明,当 $f(\boldsymbol{X})$是具有一阶连续偏导数的凸函数时,如果由最速下降法所得的点列$\{\boldsymbol{X}^{(k)}\}$有界,则必有:①数列$\{f(\boldsymbol{X}^{(k)})\}$单调下降;②序列$\{\boldsymbol{X}^{(k)}\}$的极限 $\boldsymbol{X}^*$ 满足 $\nabla f(\boldsymbol{X}^*)=0$;③$\boldsymbol{X}^*$ 为全局极小点。

由于负梯度方向的最速下降性,很容易使人们认为负梯度方向是理想的搜索方向,最速下降法是一种理想的极小化方法。必须指出,$\boldsymbol{X}$ 点处的负梯度方向 $-\nabla f(\boldsymbol{X})$,仅在 $\boldsymbol{X}$ 点附近才具有这种"最速下降"的性质,而对于整个极小化过程来说,那就是另外一回事了。由例 6 可知,若目标函数的等值线为一族同心圆(或同心球面),则从任一初始点出发,沿最速下降方向一步即可达到极小点。但通常的情况并不是这样,例如,一般二元二次凸函数的等值线为一族共心椭圆,当用最速下降法趋近极小点时,其搜索路径呈**直角锯齿状**(图 7-14)。在开头几步,目标函数值下降较快,但接近极小点 $\boldsymbol{X}^*$ 时,收敛速度就不理想了。特别是当目标函数的等值线椭圆比较扁平时,收敛速度就更慢了。因此,在实用中,常将梯度法和其他方法联合起来应用,在前期使用梯度法,而在接近极小点时,则使用收敛较快的其他方法。

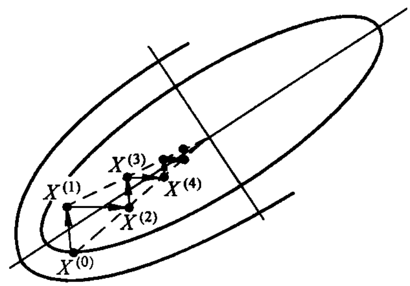


图 7-14

7.3.2 共轭梯度法

1. 共轭方向

设 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 是 n 维欧氏空间 E^n 中的两个向量,若有

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}=0$$

就称 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 正交。再设 $\boldsymbol{A}$ 为 $n\times n$ **对称正定阵**,如果 $\boldsymbol{X}$ 和 $\boldsymbol{AY}$ 正交,即有

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Y}=0 \tag{7-48}$$

则称 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 关于 $\boldsymbol{A}$ 共轭,或 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 为 $\boldsymbol{A}$ **共轭**($\boldsymbol{A}$ 正交)。

一般地,设 $\boldsymbol{A}$ 为 $n\times n$ 对称正定阵,若非零向量组 $\boldsymbol{P}^{(1)},\boldsymbol{P}^{(2)},\cdots,\boldsymbol{P}^{(n)}\in E^n$ 满足条件

$$(\boldsymbol{P}^{(i)})^{\mathrm{T}}\boldsymbol{A}\boldsymbol{P}^{(j)}=0 \quad (i\neq j;\quad i,j=1,2,\cdots,n) \tag{7-49}$$

则称该向量组为 $\boldsymbol{A}$ 共轭。如果 $\boldsymbol{A}=\boldsymbol{I}$(单位阵),则上述条件即为通常的正交条件。因此,$\boldsymbol{A}$ 共轭概念实际上是通常正交概念的推广。

定理 8 设 $\boldsymbol{A}$ 为 $n\times n$ 对称正定阵,$\boldsymbol{P}^{(1)},\boldsymbol{P}^{(2)},\cdots,\boldsymbol{P}^{(n)}$ 为 $\boldsymbol{A}$ 共轭的非零向量,则这一组向量线性独立。

证明 设向量 $\boldsymbol{P}^{(1)},\boldsymbol{P}^{(2)},\cdots,\boldsymbol{P}^{(n)}$ 之间存在如下线性关系

$$\alpha_1\boldsymbol{P}^{(1)}+\alpha_2\boldsymbol{P}^{(2)}+\cdots+\alpha_n\boldsymbol{P}^{(n)}=0$$

对 $i=1,2,\cdots,n$,用 $(\boldsymbol{P}^{(i)})^{\mathrm{T}}A$ 左乘上式得

$$\alpha_i(\boldsymbol{P}^{(i)})^{\mathrm{T}}\boldsymbol{A}\boldsymbol{P}^{(i)}=0$$

但 $\boldsymbol{P}^{(i)}\neq 0$,$\boldsymbol{A}$ 为正定,即

$$(\boldsymbol{P}^{(i)})^{\mathrm{T}}\boldsymbol{A}\boldsymbol{P}^{(i)}>0$$

故必有

$$\alpha_i=0,\quad i=1,2,\cdots,n$$

从而 $\boldsymbol{P}^{(1)},\boldsymbol{P}^{(2)},\cdots,\boldsymbol{P}^{(n)}$ 线性独立。

无约束极值问题的一个特殊情形是

$$\min f(\boldsymbol{X})=\frac{1}{2}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}+\boldsymbol{B}^{\mathrm{T}}\boldsymbol{X}+c \tag{7-50}$$

式中 $\boldsymbol{A}$ 为 $n\times n$ 对称正定阵;$\boldsymbol{X},\boldsymbol{B}\in E^n$;$c$ 为常数。问题式(7-50)称为**正定二次函数极小问题**,它在整个最优化问题中起着极其重要的作用。

定理 9 设向量 $\boldsymbol{P}^{(i)}$,$i=0,1,2,\cdots,n-1$,为 $\boldsymbol{A}$ 共轭,则从任一点 $\boldsymbol{X}^{(0)}$ 出发,相继以 $\boldsymbol{P}^{(0)},\boldsymbol{P}^{(1)},\cdots,\boldsymbol{P}^{(n-1)}$ 为搜索方向的下述算法

$$\begin{cases}\min\limits_{\lambda} f(\boldsymbol{X}^{(k)}+\lambda\boldsymbol{P}^{(k)})=f(\boldsymbol{X}^{(k)}+\lambda_k\boldsymbol{P}^{(k)})\\ \boldsymbol{X}^{(k+1)}=\boldsymbol{X}^{(k)}+\lambda_k\boldsymbol{P}^{(k)}\end{cases}$$

经 n 次一维搜索收敛于问题式(7-50)的极小点 $\boldsymbol{X}^*$。

证明 由式(7-50)

$$\nabla f(\boldsymbol{X})=\boldsymbol{A}\boldsymbol{X}+\boldsymbol{B}$$

设相继各次搜索得到的近似解分别为 $\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)},\cdots,\boldsymbol{X}^{(n)}$,则

$$\nabla f(\boldsymbol{X}^{(k)})=\boldsymbol{A}\boldsymbol{X}^{(k)}+\boldsymbol{B}$$

$$\begin{aligned}\nabla f(\boldsymbol{X}^{(k+1)})&=\boldsymbol{A}\boldsymbol{X}^{(k+1)}+\boldsymbol{B}=\boldsymbol{A}(\boldsymbol{X}^{(k)}+\lambda_k\boldsymbol{P}^{(k)})+\boldsymbol{B}\\&=\nabla f(\boldsymbol{X}^{(k)})+\lambda_k\boldsymbol{A}\boldsymbol{P}^{(k)}\end{aligned}$$

假定 $\nabla f(\boldsymbol{X}^{(k)})\neq 0$,$k=0,1,2,\cdots,n-1$,则有

$$\begin{aligned}\nabla f(\boldsymbol{X}^{(n)})&=\nabla f(\boldsymbol{X}^{(n-1)})+\lambda_{n-1}\boldsymbol{A}\boldsymbol{P}^{(n-1)}=\cdots\\&=\nabla f(\boldsymbol{X}^{(k+1)})+\lambda_{k+1}\boldsymbol{A}\boldsymbol{P}^{(k+1)}+\lambda_{k+2}\boldsymbol{A}\boldsymbol{P}^{(k+2)}+\cdots+\\&\quad\lambda_{n-1}\boldsymbol{A}\boldsymbol{P}^{(n-1)}\end{aligned}$$

由于在进行一维搜索时,为确定最佳步长 λ_k,令

$$\frac{\mathrm{d}f(\boldsymbol{X}^{(k+1)})}{\mathrm{d}\lambda}=\frac{\mathrm{d}f[\boldsymbol{X}^{(k)}+\lambda\boldsymbol{P}^{(k)}]}{\mathrm{d}\lambda}=\nabla f(\boldsymbol{X}^{(k+1)})^{\mathrm{T}}\boldsymbol{P}^{(k)}=0 \tag{7-51}$$

故对 $k=0,1,2,\cdots,n-1$ 有

$$(\boldsymbol{P}^{(k)})^{\mathrm{T}}\nabla f(\boldsymbol{X}^{(n)})=(\boldsymbol{P}^{(k)})^{\mathrm{T}}\nabla f(\boldsymbol{X}^{(k+1)})+\lambda_{k+1}(\boldsymbol{P}^{(k)})^{\mathrm{T}}\boldsymbol{A}\boldsymbol{P}^{(k+1)}+\cdots+\lambda_{n-1}(\boldsymbol{P}^{(k)})^{\mathrm{T}}\boldsymbol{A}\boldsymbol{P}^{(n-1)}=0$$

这就是 $\nabla f(\boldsymbol{X}^{(n)})$ 和 n 个线性独立的向量 $\boldsymbol{P}^{(0)},\boldsymbol{P}^{(1)},\cdots,\boldsymbol{P}^{(n-1)}$（它们为 $\boldsymbol{A}$ 共轭）正交，从而必有

$$\nabla f(\boldsymbol{X}^{(n)})=0$$

即 $\boldsymbol{X}^{(n)}$ 为 $f(\boldsymbol{X})$ 的极小点 $\boldsymbol{X}^*$。

下面我们就二维正定二次函数的情况加以说明，以便对上述定理有个直观认识。

二维正定二次函数的等值线，在极小点附近可用一族**共心椭圆**来代表（图 7-15）。大家知道，过椭圆族中心 $\boldsymbol{X}^*$ 引任意直线，必与诸椭圆相交，各交点处的切线互相平行。如果在两个互相平行的方向上进行最优一维搜索，则可得 $f(\boldsymbol{X})$ 在此方向上的极小点 $\boldsymbol{X}^{(1)}$ 和 $\bar{\boldsymbol{X}}^{(1)}$，此两点必为椭圆族中某椭圆与该平行直线的切点，而且联结 $\boldsymbol{X}^{(1)}$ 和 $\bar{\boldsymbol{X}}^{(1)}$ 的直线必通过椭圆族的中心 $\boldsymbol{X}^*$。

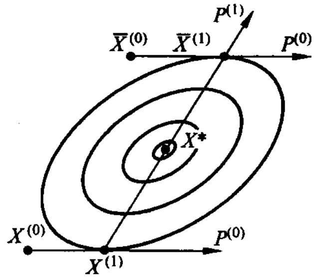


图 7-15

现从任一点 $\boldsymbol{X}^{(0)}$ 出发，沿射线 $\boldsymbol{P}^{(0)}$ 作一维搜索，则可得问题式(7-50)的目标函数 $f(\boldsymbol{X})$ 在射线 $\boldsymbol{X}^{(0)}+\lambda\boldsymbol{P}^{(0)}$ 上的极小点

$$\boldsymbol{X}^{(1)}=\boldsymbol{X}^{(0)}+\lambda_0\boldsymbol{P}^{(0)}$$

其中，λ_0 满足

$$\nabla f(\boldsymbol{X}^{(1)})^{\mathrm{T}}\boldsymbol{P}^{(0)}=0$$

同样，从另一点 $\bar{\boldsymbol{X}}^{(0)}$ 出发也沿 $\boldsymbol{P}^{(0)}$ 方向作一维搜索，可得式(7-50)中 $f(\boldsymbol{X})$ 在射线 $\bar{\boldsymbol{X}}^{(0)}+\lambda\boldsymbol{P}^{(0)}$ 上的极小点

$$\bar{\boldsymbol{X}}^{(1)}=\bar{\boldsymbol{X}}^{(0)}+\lambda_0\boldsymbol{P}^{(0)}$$

其中，λ_0 满足

$$\nabla f(\bar{\boldsymbol{X}}^{(1)})^{\mathrm{T}}\boldsymbol{P}^{(0)}=0$$

从而

$$[\nabla f(\bar{\boldsymbol{X}}^{(1)})-\nabla f(\boldsymbol{X}^{(1)})]^{\mathrm{T}}\boldsymbol{P}^{(0)}=0$$

但由式(7-50)

$$\nabla f(\boldsymbol{X})=\boldsymbol{A}\boldsymbol{X}+\boldsymbol{B}$$

若令

$$\boldsymbol{P}^{(1)}=\bar{\boldsymbol{X}}^{(1)}-\boldsymbol{X}^{(1)}$$

则有

$$(\boldsymbol{P}^{(1)})^{\mathrm{T}}\boldsymbol{A}\boldsymbol{P}^{(0)}=0$$

即 $\boldsymbol{P}^{(1)}$ 和 $\boldsymbol{P}^{(0)}$ 为 $\boldsymbol{A}$ 共轭。

上述分析说明，对于二维正定二次函数来说，从任一点 $\boldsymbol{X}^{(0)}$ 出发，沿相互共轭的方向 $\boldsymbol{P}^{(0)}$ 和 $\boldsymbol{P}^{(1)}$ 进行两次一维搜索，即可收敛到函数的极小点。

2. 正定二次函数的共轭梯度法

对于问题式(7-50)来说，由于 $\boldsymbol{A}$ 为对称正定阵，故存在唯一极小点 $\boldsymbol{X}^*$，它满足方程

$$\nabla f(\boldsymbol{X})=\boldsymbol{A}\boldsymbol{X}+\boldsymbol{B}=0 \tag{7-52}$$

且具有形式

$$\boldsymbol{X}^*=-\boldsymbol{A}^{-1}\boldsymbol{B} \tag{7-53}$$

如果已知某共轭向量组 $\boldsymbol{P}^{(0)},\boldsymbol{P}^{(1)},\cdots,\boldsymbol{P}^{(n-1)}$，由定理 9 可知，问题式(7-50)的极小点 $\boldsymbol{X}^*$ 可通过下列算法得到

$$\begin{cases}\boldsymbol{X}^{(k+1)}=\boldsymbol{X}^{(k)}+\lambda_k\boldsymbol{P}^{(k)}, \quad k=0,1,2,\cdots,n-1\\ \lambda_k:\ \min\limits_{\lambda} f(\boldsymbol{X}^{(k)}+\lambda\boldsymbol{P}^{(k)})\\ \boldsymbol{X}^{(n)}=\boldsymbol{X}^*\end{cases} \tag{7-54}$$

算法式(7-54)称为**共轭方向法**。它要求：搜索方向 $\boldsymbol{P}^{(0)},\boldsymbol{P}^{(1)},\cdots,\boldsymbol{P}^{(n-1)}$ 必须共轭；确定各近似极小点时必须按最优一维搜索进行。

共轭梯度法是共轭方向法的一种，它的搜索方向是利用一维搜索所得极小点处函数的梯度生成的，我们现在就来构造正定二次函数的**共轭梯度法**。

由于 $\nabla f(\boldsymbol{X})=\boldsymbol{A}\boldsymbol{X}+\boldsymbol{B}$，故有

$$\nabla f(\boldsymbol{X}^{(k+1)})-\nabla f(\boldsymbol{X}^{(k)})=\boldsymbol{A}(\boldsymbol{X}^{(k+1)}-\boldsymbol{X}^{(k)})$$

但

$$\boldsymbol{X}^{(k+1)}=\boldsymbol{X}^{(k)}+\lambda_k\boldsymbol{P}^{(K)}$$

故

$$\nabla f(\boldsymbol{X}^{(k+1)})-\nabla f(\boldsymbol{X}^{(k)})=\lambda_k\boldsymbol{A}\boldsymbol{P}^{(k)}, \quad k=0,1,2,\cdots,n-1 \tag{7-55}$$

任取初始近似点 $\boldsymbol{X}^{(0)}$，并取初始搜索方向为此点的负梯度方向，即

$$\boldsymbol{P}^{(0)}=-\nabla f(\boldsymbol{X}^{(0)})$$

沿射线 $\boldsymbol{X}^{(0)}+\lambda\boldsymbol{P}^{(0)}$ 进行一维搜索，得

$$\begin{cases}\boldsymbol{X}^{(1)}=\boldsymbol{X}^{(0)}+\lambda_0\boldsymbol{P}^{(0)}\\ \lambda_0:\ \min\limits_{\lambda} f(\boldsymbol{X}^{(0)}+\lambda\boldsymbol{P}^{(0)})\end{cases}$$

算出 $\nabla f(\boldsymbol{X}^{(1)})$，由式(7-51)知

$$\nabla f(\boldsymbol{X}^{(1)})^{\mathrm{T}}\boldsymbol{P}^{(0)}=-\nabla f(\boldsymbol{X}^{(1)})^{\mathrm{T}}\ \nabla f(\boldsymbol{X}^{(0)})=0$$

从而可知 $\nabla f(\boldsymbol{X}^{(1)})$ 和 $\nabla f(\boldsymbol{X}^{(0)})$ 正交(这里假设 $\nabla f(\boldsymbol{X}^{(1)})$ 和 $\nabla f(\boldsymbol{X}^{(0)})$ 均不等于零)。$\nabla f(\boldsymbol{X}^{(0)})$ 和 $\nabla f(\boldsymbol{X}^{(1)})$ 构成一正交系，我们可以在由它们生成的二维子空间中寻求 $\boldsymbol{P}^{(1)}$。

为此，可令

$$P^{(1)}=-\nabla f(X^{(1)})+\alpha_0\nabla f(X^{(0)})$$

式中，α_0 为待定系数。欲使 $P^{(1)}$ 与 $P^{(0)}$ 为 A 共轭，由式(7-55)，必须

$$[-\nabla f(X^{(1)})+\alpha_0\nabla f(X^{(0)})]^{T}[\nabla f(X^{(1)})-\nabla f(X^{(0)})]=0$$

故

$$-\alpha_0=\frac{\nabla f(X^{(1)})^{T}\nabla f(X^{(1)})}{\nabla f(X^{(0)})^{T}\nabla f(X^{(0)})}$$

令

$$\beta_0=-\alpha_0=\frac{\nabla f(X^{(1)})^{T}\nabla f(X^{(1)})}{\nabla f(X^{(0)})^{T}\nabla f(X^{(0)})} \tag{7-56}$$

由此可得

$$P^{(1)}=-\nabla f(X^{(1)})+\beta_0 P^{(0)} \tag{7-57}$$

以 $P^{(1)}$ 为搜索方向进行最优一维搜索，可得

$$\begin{cases}X^{(2)}=X^{(1)}+\lambda_1 P^{(1)}\\ \lambda_1:\ \min\limits_{\lambda} f(X^{(1)}+\lambda P^{(1)})\end{cases}$$

算出$\nabla f(X^{(2)})$，假定$\nabla f(X^{(2)})\neq 0$，因 $P^{(0)}$ 和 $P^{(1)}$ 为 A 共轭，故

$$\nabla f(X^{(0)})^{T}[\nabla f(X^{(2)})-\nabla f(X^{(1)})]=0$$

但

$$\nabla f(X^{(0)})^{T}\nabla f(X^{(1)})=0$$

故

$$\nabla f(X^{(0)})^{T}\nabla f(X^{(2)})=0$$

由于

$$\nabla f(X^{(2)})^{T}P^{(1)}=\nabla f(X^{(2)})^{T}[-\nabla f(X^{(1)})+\alpha_0\nabla f(X^{(0)})]=0$$

所以

$$\nabla f(X^{(2)})^{T}\nabla f(X^{(1)})=0$$

即$\nabla f(X^{(2)})$、$\nabla f(X^{(1)})$和$\nabla f(X^{(0)})$构成一正交系。现由它们生成的三维子空间中，寻求与 $P^{(0)}$ 和 $P^{(1)}$ 为 A 共轭的搜索方向 $P^{(2)}$。令

$$P^{(2)}=-\nabla f(X^{(2)})+\alpha_1\nabla f(X^{(1)})+\alpha_0\nabla f(X^{(0)})$$

式中，α_0 和 α_1 均为待定系数。由于 $P^{(2)}$ 应与 $P^{(0)}$ 和 $P^{(1)}$ 为 A 共轭，故须

$$[-\nabla f(X^{(2)})+\alpha_1\nabla f(X^{(1)})+\alpha_0\nabla f(X^{(0)})]^{T}[\nabla f(X^{(1)})-\nabla f(X^{(0)})]=0$$

$$[-\nabla f(X^{(2)})+\alpha_1\nabla f(X^{(1)})+\alpha_0\nabla f(X^{(0)})]^{T}[\nabla f(X^{(2)})-\nabla f(X^{(1)})]=0$$

从而

$$\alpha_1\nabla f(X^{(1)})^{T}\nabla f(X^{(1)})-\alpha_0\nabla f(X^{(0)})^{T}\nabla f(X^{(0)})=0$$

$$-\nabla f(X^{(2)})^{T}\nabla f(X^{(2)})-\alpha_1\nabla f(X^{(1)})^{T}\nabla f(X^{(1)})=0$$

解之得

$$-\alpha_1=\frac{\nabla f(\boldsymbol{X}^{(2)})^{\mathrm{T}}\ \nabla f(\boldsymbol{X}^{(2)})}{\nabla f(\boldsymbol{X}^{(1)})^{\mathrm{T}}\ \nabla f(\boldsymbol{X}^{(1)})}$$

$$\alpha_0=\alpha_1\ \frac{\nabla f(\boldsymbol{X}^{(1)})^{\mathrm{T}}\ \nabla f(\boldsymbol{X}^{(1)})}{\nabla f(\boldsymbol{X}^{(0)})^{\mathrm{T}}\ \nabla f(\boldsymbol{X}^{(0)})}$$

令 $\beta_1=-\alpha_1$，则 $\alpha_0=-\beta_1\beta_0$，于是

$$\begin{aligned}\boldsymbol{P}^{(2)}&=-\nabla f(\boldsymbol{X}^{(2)})-\beta_1\ \nabla f(\boldsymbol{X}^{(1)})-\beta_0\beta_1\ \nabla f(\boldsymbol{X}^{(0)})\\&=-\nabla f(\boldsymbol{X}^{(2)})+\beta_1[-\nabla f(\boldsymbol{X}^{(1)})-\beta_0\ \nabla f(\boldsymbol{X}^{(0)})]\\&=-\nabla f(\boldsymbol{X}^{(2)})+\beta_1[-\nabla f(\boldsymbol{X}^{(1)})+\beta_0\boldsymbol{P}^{(0)}]\\&=-\nabla f(\boldsymbol{X}^{(2)})+\beta_1\boldsymbol{P}^{(1)}\end{aligned}\tag{7-58}$$

继续上述步骤，可得一般公式如下

$$\begin{cases}\boldsymbol{P}^{(k+1)}=-\nabla f(\boldsymbol{X}^{(k+1)})+\beta_k\boldsymbol{P}^{(k)}\\ \beta_k=\dfrac{\nabla f(\boldsymbol{X}^{(k+1)})^{\mathrm{T}}\ \nabla f(\boldsymbol{X}^{(k+1)})}{\nabla f(\boldsymbol{X}^{(k)})^{\mathrm{T}}\ \nabla f(\boldsymbol{X}^{(k)})}\end{cases}$$

对于正定二次函数来说，$\nabla f(\boldsymbol{X})=\boldsymbol{AX}+\boldsymbol{B}$，由式(7-55)

$$\nabla f(\boldsymbol{X}^{(k+1)})=\nabla f(\boldsymbol{X}^{(k)})+\lambda_k\boldsymbol{AP}^{(k)}$$

由于进行的是最优一维搜索，故有

$$\nabla f(\boldsymbol{X}^{(k+1)})^{\mathrm{T}}\boldsymbol{P}^{(k)}=0$$

从而

$$\lambda_k=-\frac{\nabla f(\boldsymbol{X}^{(k)})^{\mathrm{T}}\boldsymbol{P}^{(k)}}{(\boldsymbol{P}^{(k)})^{\mathrm{T}}\boldsymbol{AP}^{(k)}}$$

如此，即可得共轭梯度法的一组计算公式如下

$$\begin{cases}\boldsymbol{X}^{(k+1)}=\boldsymbol{X}^{(k)}+\lambda_k\boldsymbol{P}^{(k)} & (7\text{-}59)\\ \lambda_k=-\dfrac{\nabla f(\boldsymbol{X}^{(k)})^{\mathrm{T}}\boldsymbol{P}^{(k)}}{(\boldsymbol{P}^{(k)})^{\mathrm{T}}\boldsymbol{AP}^{(k)}} & (7\text{-}60)\\ \boldsymbol{P}^{(k+1)}=-\nabla f(\boldsymbol{X}^{(k+1)})+\beta_k\boldsymbol{P}^{(k)} & (7\text{-}61)\\ \beta_k=\dfrac{\nabla f(\boldsymbol{X}^{(k+1)})^{\mathrm{T}}\ \nabla f(\boldsymbol{X}^{(k+1)})}{\nabla f(\boldsymbol{X}^{(k)})^{\mathrm{T}}\ \nabla f(\boldsymbol{X}^{(k)})} & (7\text{-}62)\\ \qquad k=0,1,2,\cdots,n-1\end{cases}$$

其中，$\boldsymbol{X}^{(0)}$ 为初始近似，$\boldsymbol{P}^{(0)}=-\nabla f(\boldsymbol{X}^{(0)})$。

由于 $\boldsymbol{P}^{(k)}=-\nabla f(\boldsymbol{X}^{(k)})+\beta_{k-1}\boldsymbol{P}^{(k-1)}$ 以及 $\nabla f(\boldsymbol{X}^{(k)})^{\mathrm{T}}\boldsymbol{P}^{(k-1)}=0$，故式(7-60)也可写成

$$\lambda_k=\frac{\nabla f(\boldsymbol{X}^{(k)})^{\mathrm{T}}\ \nabla f(\boldsymbol{X}^{(k)})}{(\boldsymbol{P}^{(k)})^{\mathrm{T}}\boldsymbol{AP}^{(k)}}\tag{7-63}$$

式(7-62)最先由弗莱彻(Fletcher)和瑞夫斯(Reeves)提出，故此法亦称为 FR 共轭梯度法。上述公式还有其他等价形式。例如，借助于式(7-55)，可将它变为

$$\beta_k=\frac{\nabla f(\boldsymbol{X}^{(k+1)})^{\mathrm{T}}\boldsymbol{AP}^{(k)}}{(\boldsymbol{P}^{(k)})^{\mathrm{T}}\boldsymbol{AP}^{(k)}}\tag{7-64}$$

现将共轭梯度法的计算步骤总结如下：

(1) 选择初始近似 $\boldsymbol{X}^{(0)}$，给出允许误差 $\varepsilon>0$。

(2) 计算

$$\boldsymbol{P}^{(0)}=-\nabla f(\boldsymbol{X}^{(0)})$$

并用式(7-59)和式(7-60)算出 $\boldsymbol{X}^{(1)}$。计算步长也可使用以前介绍过的一维搜索法。

(3) 一般地，假定已得出 $\boldsymbol{X}^{(k)}$ 和 $\boldsymbol{P}^{(k)}$，则可计算其第 $k+1$ 次近似 $\boldsymbol{X}^{(k+1)}$

$$\begin{cases}\boldsymbol{X}^{(k+1)}=\boldsymbol{X}^{(k)}+\lambda_k\boldsymbol{P}^{(k)}\\ \lambda_k:\ \min\limits_{\lambda} f(\boldsymbol{X}^{(k)}+\lambda\boldsymbol{P}^{(k)})\end{cases}$$

(4) 若 $\|\nabla f(\boldsymbol{X}^{(k+1)})\|^2\leqslant\varepsilon$，停止计算，$\boldsymbol{X}^{(k+1)}$ 即为要求的近似解。否则，若 $k<n-1$，则用式(7-62)和式(7-61)计算 β_k 和 $\boldsymbol{P}^{(k+1)}$，并转向第 3 步。

应当指出，对于二次函数的情形，从理论上说，进行 n 次迭代即可达到极小点。但是，在实际计算中，由于数据的四舍五入以及计算误差的积累，往往做不到这一点。此外，由于 n 维问题的共轭方向最多只有 n 个，在 n 步以后继续如上进行是没有意义的。因此，在实际应用时，如迭代到 n 步还不收敛，就将 $\boldsymbol{X}^{(n)}$ 作为新的初始近似，重新开始迭代。根据实际经验，采用这种再开始的办法，一般都可得到较好的效果。

例 8 试用共轭梯度法求下述二次函数的极小点

$$f(\boldsymbol{X})=\frac{3}{2}x_1^2+\frac{1}{2}x_2^2-x_1x_2-2x_1$$

解 将 $f(\boldsymbol{X})$ 化成式(7-50)的形式，得

$$\boldsymbol{A}=\begin{pmatrix}3 & -1\\ -1 & 1\end{pmatrix}$$

现从 $\boldsymbol{X}^{(0)}=(-2,4)^{\mathrm{T}}$ 开始，由于

$$\nabla f(\boldsymbol{X})=[(3x_1-x_2-2),(x_2-x_1)]^{\mathrm{T}}$$

故

$$\nabla f(\boldsymbol{X}^{(0)})=(-12,6)^{\mathrm{T}}$$

$$\boldsymbol{P}^{(0)}=-\nabla f(\boldsymbol{X}^{(0)})=(12,-6)^{\mathrm{T}}$$

$$\lambda_0=-\frac{\nabla f(\boldsymbol{X}^{(0)})^{\mathrm{T}}\boldsymbol{P}^{(0)}}{(\boldsymbol{P}^{(0)})^{\mathrm{T}}\boldsymbol{A}\boldsymbol{P}^{(0)}}=-\frac{(-12,6)\begin{pmatrix}12\\ -6\end{pmatrix}}{(12,-6)\begin{pmatrix}3 & -1\\ -1 & 1\end{pmatrix}\begin{pmatrix}12\\ -6\end{pmatrix}}=\frac{180}{612}=\frac{5}{17}$$

于是

$$\boldsymbol{X}^{(1)}=\boldsymbol{X}^{(0)}+\lambda_0\boldsymbol{P}^{(0)}=\begin{bmatrix}-2\\ 4\end{bmatrix}+\frac{5}{17}\begin{bmatrix}12\\ -6\end{bmatrix}=\left(\frac{26}{17},\frac{38}{17}\right)^{\mathrm{T}}$$

$$\nabla f(\boldsymbol{X}^{(1)})=(6/17,12/17)^{\mathrm{T}}$$

$$\beta_0=\frac{\nabla f(\boldsymbol{X}^{(1)})^{\mathrm{T}}\ \nabla f(\boldsymbol{X}^{(1)})}{\nabla f(\boldsymbol{X}^{(0)})^{\mathrm{T}}\ \nabla f(\boldsymbol{X}^{(0)})}=\frac{\left(\frac{6}{17},\frac{12}{17}\right)\begin{bmatrix}6/17\\12/17\end{bmatrix}}{(-12,6)\begin{pmatrix}-12\\6\end{pmatrix}}=\frac{1}{289}$$

$$\boldsymbol{P}^{(1)}=-\nabla f(\boldsymbol{X}^{(1)})+\beta_0\boldsymbol{P}^{(0)}=-\begin{bmatrix}6/17\\12/17\end{bmatrix}+\frac{1}{289}\begin{pmatrix}12\\-6\end{pmatrix}=\left(-\frac{90}{289},-\frac{210}{289}\right)^{\mathrm{T}}$$

$$\begin{aligned}\lambda_1&=-\frac{\nabla f(\boldsymbol{X}^{(1)})^{\mathrm{T}}\boldsymbol{P}^{(1)}}{(\boldsymbol{P}^{(1)})^{\mathrm{T}}\boldsymbol{A}\boldsymbol{P}^{(1)}}\\&=-\frac{\left(\frac{6}{17},\frac{12}{17}\right)\left(-\frac{90}{289},-\frac{210}{289}\right)^{\mathrm{T}}}{\left(-\frac{90}{289},-\frac{210}{289}\right)\begin{bmatrix}3&-1\\-1&1\end{bmatrix}\left(-\frac{90}{289},-\frac{210}{289}\right)^{\mathrm{T}}}\\&=\frac{6\times17\times90+12\times17\times210}{(-60,-120)(-90,-210)^{\mathrm{T}}}=\frac{17(6\times90+12\times210)}{60\times90+120\times210}\\&=\frac{17}{10}\end{aligned}$$

故

$$\boldsymbol{X}^{(2)}=\boldsymbol{X}^{(1)}+\lambda_1\boldsymbol{P}^{(1)}=\begin{bmatrix}26/17\\38/17\end{bmatrix}+\frac{17}{10}\begin{bmatrix}-90/289\\-210/289\end{bmatrix}=\begin{pmatrix}1\\1\end{pmatrix}$$

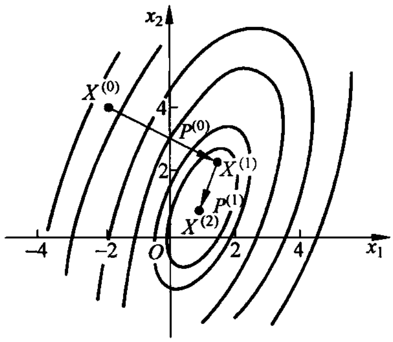


图 7-16

这就是 $f(\boldsymbol{X})$ 的极小点。图 7-16 表明了本例的搜索方向和步骤。

3. 非二次函数的共轭梯度法

可将共轭梯度法推广到求解一般无约束极值问题式(7-40)。

设 $f(\boldsymbol{X})$ 为某一严格凸函数，它具有二阶连续偏导数，其唯一极小点为 $\boldsymbol{X}^*$。现任取初始近似 $\boldsymbol{X}^{(0)}$，计算 $\nabla f(\boldsymbol{X}^{(0)})$，选取 $\boldsymbol{P}^{(0)}=-\nabla f(\boldsymbol{X}^{(0)})$ 为初始搜索方向，作射线 $\boldsymbol{X}^{(0)}+\lambda\boldsymbol{P}^{(0)}$ $(\lambda\geqslant 0)$，并将 $f(\boldsymbol{X})=f(\boldsymbol{X}^{(0)}+\lambda\boldsymbol{P}^{(0)})$ 于 $\boldsymbol{X}^{(0)}$ 附近做泰勒展开

$$f(\boldsymbol{X}^{(0)}+\lambda\boldsymbol{P}^{(0)})\approx f(\boldsymbol{X}^{(0)})+\lambda\ \nabla f(\boldsymbol{X}^{(0)})^{\mathrm{T}}\boldsymbol{P}^{(0)}+\frac{1}{2}\lambda^2(\boldsymbol{P}^{(0)})^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{X}^{(0)})\boldsymbol{P}^{(0)}$$

上式为 λ 的二次函数，因 $(\boldsymbol{P}^{(0)})^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{X}^{(0)})\boldsymbol{P}^{(0)}>0$，故使该二次函数沿 $\boldsymbol{P}^{(0)}$ 方向取极小值的 λ 为

$$\lambda_0=-\frac{\nabla f(\boldsymbol{X}^{(0)})^{\mathrm{T}}\boldsymbol{P}^{(0)}}{(\boldsymbol{P}^{(0)})^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{X}^{(0)})\boldsymbol{P}^{(0)}}$$

显然，它近似满足 $\min_{\lambda} f(\boldsymbol{X}^{(0)}+\lambda\boldsymbol{P}^{(0)})$。令

$$\boldsymbol{X}^{(1)}=\boldsymbol{X}^{(0)}+\lambda_0\boldsymbol{P}^{(0)}$$

则 $\boldsymbol{X}^{(1)}$ 近似满足

$$\nabla f(\boldsymbol{X}^{(1)})^{\mathrm{T}}\boldsymbol{P}^{(0)}=0$$

现构造向量

$$\boldsymbol{P}^{(1)}=-\nabla f(\boldsymbol{X}^{(1)})+\beta_0\boldsymbol{P}^{(0)}$$

使满足 $(\boldsymbol{P}^{(1)})^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{X}^{(0)})\boldsymbol{P}^{(0)}=0$，则得

$$\beta_0=\frac{\nabla f(\boldsymbol{X}^{(1)})^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{X}^{(0)})\boldsymbol{P}^{(0)}}{(\boldsymbol{P}^{(0)})^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{X}^{(0)})\boldsymbol{P}^{(0)}}$$

这就确定了 $\boldsymbol{P}^{(1)}$。

按此手续可构造各次迭代的搜索方向及近似点。一般地，我们有

$$\begin{cases}\boldsymbol{X}^{(k+1)}=\boldsymbol{X}^{(k)}+\lambda_k\boldsymbol{P}^{(k)} \\ \lambda_k=-\dfrac{\nabla f(\boldsymbol{X}^{(k)})^{\mathrm{T}}\boldsymbol{P}^{(k)}}{(\boldsymbol{P}^{(k)})^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{X}^{(k)})\boldsymbol{P}^{(k)}} & (7\text{-}65) \\ \boldsymbol{P}^{(k+1)}=-\nabla f(\boldsymbol{X}^{(k+1)})+\beta_k\boldsymbol{P}^{(k)} \\ \beta_k=\dfrac{\nabla f(\boldsymbol{X}^{(k+1)})^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{X}^{(k)})\boldsymbol{P}^{(k)}}{(\boldsymbol{P}^{(k)})^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{X}^{(k)})\boldsymbol{P}^{(k)}} & (7\text{-}66)\end{cases}$$

这就是推广到非二次函数的共轭梯度法的计算公式。

由于在导出上述公式的过程中利用了一些近似关系，以及 $\boldsymbol{H}(\boldsymbol{X}^{(k)})$ 的逐次变化，使 $\boldsymbol{P}^{(0)},\boldsymbol{P}^{(1)},\cdots,\boldsymbol{P}^{(n-1)}$ 的共轭性遭受破坏，因而对于一般非二次函数来说，要以 n 步迭代取得收敛常常是不可能的。所以在实际应用时，如迭代步数 $k\leqslant n$ 已达到要求的精度，则以 $\boldsymbol{X}^{(k)}$ 作为要求的近似解。否则可将前 n 步作为一个循环，同时以所得到的 $\boldsymbol{X}^{(n)}$ 作为新的初始近似重新开始，进行第二个循环。重复进行，直至满足要求的精度为止。

7.3.3　变尺度法

变尺度法是近 40 多年来发展起来的，它是求解无约束极值问题的一种有效方法。由于它既避免了计算二阶导数矩阵及其求逆过程，又比梯度法的收敛速度快，特别是对高维问题具有显著的优越性，因而使变尺度法获得了很高的声誉，至今仍被公认为求解无约束极值问题最有效的算法之一。下面我们就来简要地介绍变尺度法的基本原理及其计算过程。

1. 基本原理

假定无约束极值问题的目标函数 $f(\boldsymbol{X})$ 具有二阶连续偏导数，$\boldsymbol{X}^{(k)}$ 为其极小点的某一近似。在这个点附近取 $f(\boldsymbol{X})$ 的二阶泰勒多项式逼近

$$f(\boldsymbol{X})\approx f(\boldsymbol{X}^{(k)})+\nabla f(\boldsymbol{X}^{(k)})^{\mathrm{T}}\Delta\boldsymbol{X}+\frac{1}{2}\Delta\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{X}^{(k)})\Delta\boldsymbol{X} \tag{7-67}$$

则其梯度为

$$\nabla f(\boldsymbol{X}) \approx \nabla f(\boldsymbol{X}^{(k)}) + \boldsymbol{H}(\boldsymbol{X}^{(k)})\Delta\boldsymbol{X} \tag{7-68}$$

这个近似函数的极小点满足

$$\nabla f(\boldsymbol{X}^{(k)}) + \boldsymbol{H}(\boldsymbol{X}^{(k)})\Delta\boldsymbol{X} = 0$$

从而

$$\boldsymbol{X} = \boldsymbol{X}^{(k)} - \boldsymbol{H}(\boldsymbol{X}^{(k)})^{-1}\nabla f(\boldsymbol{X}^{(k)}) \tag{7-69}$$

其中 $\boldsymbol{H}(\boldsymbol{X}^{(k)})$为 $f(\boldsymbol{X})$在 $\boldsymbol{X}^{(k)}$点的海赛矩阵。

如果 $f(\boldsymbol{X})$是二次函数，则 $\boldsymbol{H}(\boldsymbol{X})$为常数阵。这时，逼近式(7-67)是准确的。在这种情况下，从任一点 $\boldsymbol{X}^{(k)}$出发，用式(7-69)只要一步即可求出 $f(\boldsymbol{X})$的极小点(假定 $\boldsymbol{H}(\boldsymbol{X}^{(k)})$正定)。

当 $f(\boldsymbol{X})$不是二次函数时，式(7-67)仅是 $f(\boldsymbol{X})$在 $\boldsymbol{X}^{(k)}$点附近的近似表达式。这时，按式(7-69)求得的极小点，只是 $f(\boldsymbol{X})$的极小点的近似。在这种情况下，人们常取 $-\boldsymbol{H}(\boldsymbol{X}^{(k)})^{-1}\nabla f(\boldsymbol{X}^{(k)})$为搜索方向，即

$$\begin{cases} \boldsymbol{P}^{(k)} = -\boldsymbol{H}(\boldsymbol{X}^{(k)})^{-1}\nabla f(\boldsymbol{X}^{(k)}) \\ \boldsymbol{X}^{(k+1)} = \boldsymbol{X}^{(k)} + \lambda_k \boldsymbol{P}^{(k)} \\ \lambda_k:\ \min\limits_{\lambda} f(\boldsymbol{X}^{(k)} + \lambda\boldsymbol{P}^{(k)}) \end{cases} \tag{7-70}$$

按照这种方式求函数 $f(\boldsymbol{X})$的极小点的方法，称做**广义牛顿法**。式(7-70)确定的搜索方向，为 $f(\boldsymbol{X})$在点 $\boldsymbol{X}^{(k)}$的**牛顿方向**。牛顿法的收敛速度很快，当 $f(\boldsymbol{X})$的二阶导数及其海赛矩阵的逆阵便于计算时，使用这一方法非常有效。

问题在于，实际问题中的目标函数往往相当复杂，计算二阶导数的工作量或者太大，或者根本不可能。况且，在 $\boldsymbol{X}$ 的维数很高时，计算逆阵也相当费事。为了不计算二阶导数矩阵 $\boldsymbol{H}(\boldsymbol{X}^{(k)})$，从而也不必计算其逆阵 $\boldsymbol{H}(\boldsymbol{X}^{(k)})^{-1}$，我们设法构造另一个矩阵 $\overline{\boldsymbol{H}}^{(k)}$，用它来直接逼近二阶导数矩阵的逆阵 $\boldsymbol{H}(\boldsymbol{X}^{(k)})^{-1}$。

下面就来研究如何构造 $\boldsymbol{H}(\boldsymbol{X}^{(k)})^{-1}$的近似矩阵 $\overline{\boldsymbol{H}}^{(k)}$。我们要求，在每一步都能以现有的信息来确定下一个搜索方向；每做一次迭代，目标函数值均有所下降；而且，这些近似矩阵最后应收敛于解点处的海赛矩阵的逆阵。

当 $f(\boldsymbol{X})$是二次函数时，其海赛矩阵为常数阵，可知其在任两点 $\boldsymbol{X}^{(k)}$和 $\boldsymbol{X}^{(k+1)}$处的梯度之差等于

$$\nabla f(\boldsymbol{X}^{(k+1)}) - \nabla f(\boldsymbol{X}^{(k)}) = \boldsymbol{A}(\boldsymbol{X}^{(k+1)} - \boldsymbol{X}^{(k)})$$

或

$$\boldsymbol{X}^{(k+1)} - \boldsymbol{X}^{(k)} = \boldsymbol{A}^{-1}[\nabla f(\boldsymbol{X}^{(k+1)}) - \nabla f(\boldsymbol{X}^{(k)})] \tag{7-71}$$

对于非二次函数，仿照二次函数的情形，要求其海赛矩阵的逆阵的第 $k+1$ 次近似矩阵 $\overline{\boldsymbol{H}}^{(k+1)}$满足关系式

$$\boldsymbol{X}^{(k+1)} - \boldsymbol{X}^{(k)} = \overline{\boldsymbol{H}}^{(k+1)}[\nabla f(\boldsymbol{X}^{(k+1)}) - \nabla f(\boldsymbol{X}^{(k)})] \tag{7-72}$$

此式就是所谓的**拟牛顿条件**。

若令

$$\begin{cases}\Delta \boldsymbol{G}^{(k)}=\nabla f(\boldsymbol{X}^{(k+1)})-\nabla f(\boldsymbol{X}^{(k)})\\ \Delta \boldsymbol{X}^{(k)}=\boldsymbol{X}^{(k+1)}-\boldsymbol{X}^{(k)}\end{cases} \tag{7-73}$$

则式(7-72)变为

$$\Delta \boldsymbol{X}^{(k)}=\overline{\boldsymbol{H}}^{(k+1)}\Delta \boldsymbol{G}^{(k)}$$

现设 $\overline{\boldsymbol{H}}^{(k)}$ 已知,并用下式求 $\overline{\boldsymbol{H}}^{(k+1)}$(假定 $\overline{\boldsymbol{H}}^{(k)}$ 和 $\overline{\boldsymbol{H}}^{(k+1)}$ 都为对称正定阵)

$$\overline{\boldsymbol{H}}^{(k+1)}=\overline{\boldsymbol{H}}^{(k)}+\Delta \overline{\boldsymbol{H}}^{(k)} \tag{7-74}$$

上式中 $\Delta\overline{\boldsymbol{H}}^{(k)}$ 为第 k 次校正矩阵,$\overline{\boldsymbol{H}}^{(k+1)}$ 应满足拟牛顿条件,即要求

$$\Delta \boldsymbol{X}^{(k)}=(\overline{\boldsymbol{H}}^{(k)}+\Delta\overline{\boldsymbol{H}}^{(k)})\Delta \boldsymbol{G}^{(k)}$$

或

$$\Delta\overline{\boldsymbol{H}}^{(k)}\Delta \boldsymbol{G}^{(k)}=\Delta \boldsymbol{X}^{(k)}-\overline{\boldsymbol{H}}^{(k)}\Delta \boldsymbol{G}^{(k)} \tag{7-75}$$

由此可以设想 $\Delta\overline{\boldsymbol{H}}^{(k)}$ 的一种较简单形式为

$$\Delta\overline{\boldsymbol{H}}^{(k)}=\Delta \boldsymbol{X}^{(k)}(\boldsymbol{Q}^{(k)})^{\mathrm{T}}-\overline{\boldsymbol{H}}^{(k)}\Delta \boldsymbol{G}^{(k)}(\boldsymbol{W}^{(k)})^{\mathrm{T}} \tag{7-76}$$

式中 $\boldsymbol{Q}^{(k)}$ 和 $\boldsymbol{W}^{(k)}$ 为两个待定向量。

将表达式(7-76)代入式(7-75)得

$$\Delta \boldsymbol{X}^{(k)}(\boldsymbol{Q}^{(k)})^{\mathrm{T}}\Delta \boldsymbol{G}^{(k)}-\overline{\boldsymbol{H}}^{(k)}\Delta \boldsymbol{G}^{(k)}(\boldsymbol{W}^{(k)})^{\mathrm{T}}\Delta \boldsymbol{G}^{(k)}=\Delta \boldsymbol{X}^{(k)}-\overline{\boldsymbol{H}}^{(k)}\Delta \boldsymbol{G}^{(k)}$$

这就是说,应使

$$(\boldsymbol{Q}^{(k)})^{\mathrm{T}}\Delta \boldsymbol{G}^{(k)}=(\boldsymbol{W}^{(k)})^{\mathrm{T}}\Delta \boldsymbol{G}^{(k)}=1 \tag{7-77}$$

由于 $\Delta\overline{\boldsymbol{H}}^{(k)}$ 应为对称阵,最简单的办法就是取

$$\begin{cases}\boldsymbol{Q}^{(k)}=\eta_k\Delta \boldsymbol{X}^{(k)}\\ \boldsymbol{W}^{(k)}=\xi_k\overline{\boldsymbol{H}}^{(k)}\Delta \boldsymbol{G}^{(k)}\end{cases} \tag{7-78}$$

由式(7-77)

$$\eta_k(\Delta \boldsymbol{X}^{(k)})^{\mathrm{T}}\Delta \boldsymbol{G}^{(k)}=\xi_k(\Delta \boldsymbol{G}^{(k)})^{\mathrm{T}}\overline{\boldsymbol{H}}^{(k)}\Delta \boldsymbol{G}^{(k)}=1$$

设 $(\Delta \boldsymbol{X}^{(k)})^{\mathrm{T}}\Delta \boldsymbol{G}^{(k)}$ 以及 $(\Delta \boldsymbol{G}^{(k)})^{\mathrm{T}}\overline{\boldsymbol{H}}^{(k)}\Delta \boldsymbol{G}^{(k)}$ 皆不为零,则有

$$\begin{cases}\eta_k=\dfrac{1}{(\Delta \boldsymbol{X}^{(k)})^{\mathrm{T}}\Delta \boldsymbol{G}^{(K)}}=\dfrac{1}{(\Delta \boldsymbol{G}^{(k)})^{\mathrm{T}}\Delta \boldsymbol{X}^{(k)}}\\ \xi_k=\dfrac{1}{(\Delta \boldsymbol{G}^{(k)})^{\mathrm{T}}\overline{\boldsymbol{H}}^{(k)}\Delta \boldsymbol{G}^{(k)}}\end{cases} \tag{7-79}$$

于是得到校正矩阵

$$\Delta\overline{\boldsymbol{H}}^{(k)}=\frac{\Delta \boldsymbol{X}^{(k)}(\Delta \boldsymbol{X}^{(k)})^{\mathrm{T}}}{(\Delta \boldsymbol{G}^{(k)})^{\mathrm{T}}\Delta \boldsymbol{X}^{(k)}}-\frac{\overline{\boldsymbol{H}}^{(k)}\Delta \boldsymbol{G}^{(k)}(\Delta \boldsymbol{G}^{(k)})^{\mathrm{T}}\overline{\boldsymbol{H}}^{(k)}}{(\Delta \boldsymbol{G}^{(k)})^{\mathrm{T}}\overline{\boldsymbol{H}}^{(k)}\Delta \boldsymbol{G}^{(k)}} \tag{7-80}$$

从而得到

$$\overline{\boldsymbol{H}}^{(k+1)}=\overline{\boldsymbol{H}}^{(k)}+\frac{\Delta \boldsymbol{X}^{(k)}(\Delta \boldsymbol{X}^{(k)})^{\mathrm{T}}}{(\Delta \boldsymbol{G}^{(k)})^{\mathrm{T}}\Delta \boldsymbol{X}^{(k)}}-\frac{\overline{\boldsymbol{H}}^{(k)}\Delta \boldsymbol{G}^{(k)}(\Delta \boldsymbol{G}^{(k)})^{\mathrm{T}}\overline{\boldsymbol{H}}^{(k)}}{(\Delta \boldsymbol{G}^{(k)})^{\mathrm{T}}\overline{\boldsymbol{H}}^{(k)}\Delta \boldsymbol{G}^{(k)}} \tag{7-81}$$

上述矩阵称为**尺度矩阵**,在整个迭代过程中它是在不断变化的。有了尺度矩阵,即可依式(7-70)进行迭代计算。

2. 计算步骤

现将变尺度法的计算步骤总结如下：

(1) 给定初始点 $\boldsymbol{X}^{(0)}$ 及梯度允许误差 $\varepsilon>0$。

(2) 若

$$\|\nabla f(\boldsymbol{X}^{(0)})\|^2\leqslant\varepsilon$$

则 $\boldsymbol{X}^{(0)}$ 即为近似极小点，停止迭代。否则，转向下一步。

(3) 令

$$\overline{\boldsymbol{H}}^{(0)}=\boldsymbol{I}(\text{单位阵})$$

$$\boldsymbol{P}^{(0)}=-\overline{\boldsymbol{H}}^{(0)}\nabla f(\boldsymbol{X}^{(0)})$$

在 $\boldsymbol{P}^{(0)}$ 方向进行一维搜索，确定最佳步长 λ_0

$$\min_{\lambda} f(\boldsymbol{X}^{(0)}+\lambda\boldsymbol{P}^{(0)})=f(\boldsymbol{X}^{(0)}+\lambda_0\boldsymbol{P}^{(0)})$$

如此可得下一个近似点

$$\boldsymbol{X}^{(1)}=\boldsymbol{X}^{(0)}+\lambda_0\boldsymbol{P}^{(0)}$$

(4) 一般地，设已得到近似点 $\boldsymbol{X}^{(k)}$，算出 $\nabla f(\boldsymbol{X}^{(k)})$，若

$$\|\nabla f(\boldsymbol{X}^{(k)})\|^2\leqslant\varepsilon$$

则 $\boldsymbol{X}^{(k)}$ 即为所求的近似解，停止迭代；否则，按式(7-81)计算 $\overline{\boldsymbol{H}}^{(k)}$，并令

$$\boldsymbol{P}^{(k)}=-\overline{\boldsymbol{H}}^{(k)}\nabla f(\boldsymbol{X}^{(k)})$$

在 $\boldsymbol{P}^{(k)}$ 方向进行一维搜索，确定最佳步长 λ_k

$$\min_{\lambda} f(\boldsymbol{X}^{(k)}+\lambda\boldsymbol{P}^{(k)})=f(\boldsymbol{X}^{(k)}+\lambda_k\boldsymbol{P}^{(k)})$$

其下一个近似点为

$$\boldsymbol{X}^{(k+1)}=\boldsymbol{X}^{(k)}+\lambda_k\boldsymbol{P}^{(k)}$$

(5) 若 $\boldsymbol{X}^{(k+1)}$ 点满足精度要求，则 $\boldsymbol{X}^{(k+1)}$ 即为所求的近似解。否则，转回第(4)步，直到求出某点满足精度要求为止。

和共轭梯度法相类似，如果迭代 n 次仍不收敛，则以 $\boldsymbol{X}^{(n)}$ 为新的 $\boldsymbol{X}^{(0)}$，以这时的 $\boldsymbol{X}^{(0)}$ 为起点重新开始一轮新的迭代。

上述方法首先由戴维顿(Davidon)提出，后经弗莱彻(Fletcher)和鲍威尔(Powell)加以改进，故称 DFP 法，或 DFP 变尺度法。

例 9 试用 DFP 法重新计算例 8。

解 和例 8 一样，仍从 $\boldsymbol{X}^{(0)}=(-2,4)^{\mathrm{T}}$ 开始，并取

$$\overline{\boldsymbol{H}}^{(0)}=\begin{pmatrix}1&0\\0&1\end{pmatrix}$$

$$\nabla f(\boldsymbol{X})=[(3x_1-x_2-2),(x_2-x_1)]^{\mathrm{T}}$$

$$\nabla f(\boldsymbol{X}^{(0)})=(-12,6)^{\mathrm{T}}$$

$$\boldsymbol{P}^{(0)}=-\overline{\boldsymbol{H}}^{(0)}\nabla f(\boldsymbol{X}^{(0)})=-\begin{pmatrix}1&0\\0&1\end{pmatrix}\begin{pmatrix}-12\\6\end{pmatrix}=\begin{pmatrix}12\\-6\end{pmatrix}$$

利用一维搜索，即 $\min\limits_{\lambda} f(\boldsymbol{X}^{(0)}+\lambda \boldsymbol{P}^{(0)})$，可算得

$$\lambda_0=\frac{5}{17}$$

$$\boldsymbol{X}^{(1)}=\boldsymbol{X}^{(0)}+\lambda_0\boldsymbol{P}^{(0)}=\begin{pmatrix}-2\\4\end{pmatrix}+\frac{5}{17}\begin{pmatrix}12\\-6\end{pmatrix}=\left(\frac{26}{17},\frac{38}{17}\right)^{\mathrm{T}}$$

$$\nabla f(\boldsymbol{X}^{(1)})=\left(\frac{6}{17},\frac{12}{17}\right)^{\mathrm{T}}$$

$$\Delta\boldsymbol{X}^{(0)}=\boldsymbol{X}^{(1)}-\boldsymbol{X}^{(0)}=\left(\frac{26}{17},\frac{38}{17}\right)^{\mathrm{T}}-(-2,4)^{\mathrm{T}}=\left(\frac{60}{17},-\frac{30}{17}\right)^{\mathrm{T}}$$

$$\begin{aligned}\Delta\boldsymbol{G}^{(0)}&=\nabla f(\boldsymbol{X}^{(1)})-\nabla f(\boldsymbol{X}^{(0)})\\&=\left(\frac{6}{17},\frac{12}{17}\right)^{\mathrm{T}}-(-12,6)^{\mathrm{T}}=\left(\frac{210}{17},-\frac{90}{17}\right)^{\mathrm{T}}\end{aligned}$$

$$\begin{aligned}\overline{\boldsymbol{H}}^{(1)}&=\overline{\boldsymbol{H}}^{(0)}+\frac{\Delta\boldsymbol{X}^{(0)}(\Delta\boldsymbol{X}^{(0)})^{\mathrm{T}}}{(\Delta\boldsymbol{G}^{(0)})^{\mathrm{T}}\Delta\boldsymbol{X}^{(0)}}-\frac{\overline{\boldsymbol{H}}^{(0)}\Delta\boldsymbol{G}^{(0)}(\Delta\boldsymbol{G}^{(0)})^{\mathrm{T}}\overline{\boldsymbol{H}}^{(0)}}{(\Delta\boldsymbol{G}^{(0)})^{\mathrm{T}}\overline{\boldsymbol{H}}^{(0)}\Delta\boldsymbol{G}^{(0)}}\\
&=\begin{pmatrix}1&0\\0&1\end{pmatrix}+\frac{\left(\frac{60}{17},-\frac{30}{17}\right)^{\mathrm{T}}\left(\frac{60}{17},-\frac{30}{17}\right)}{\left(\frac{210}{17},-\frac{90}{17}\right)\left(\frac{60}{17},-\frac{30}{17}\right)^{\mathrm{T}}}-\\
&\quad\frac{\begin{pmatrix}1&0\\0&1\end{pmatrix}\left(\frac{210}{17},-\frac{90}{17}\right)^{\mathrm{T}}\left(\frac{210}{17},-\frac{90}{17}\right)\begin{pmatrix}1&0\\0&1\end{pmatrix}}{\left(\frac{210}{17},-\frac{90}{17}\right)\begin{pmatrix}1&0\\0&1\end{pmatrix}\left(\frac{210}{17},-\frac{90}{17}\right)^{\mathrm{T}}}\\
&=\begin{pmatrix}1&0\\0&1\end{pmatrix}+\frac{1}{17}\begin{pmatrix}4&-2\\-2&1\end{pmatrix}-\frac{1}{58}\begin{pmatrix}49&-21\\-21&9\end{pmatrix}\\
&=\frac{1}{986}\begin{pmatrix}385&241\\241&891\end{pmatrix}\end{aligned}$$

$$\boldsymbol{P}^{(1)}=-\overline{\boldsymbol{H}}^{(1)}\nabla f(\boldsymbol{X}^{(1)})=-\frac{1}{986}\begin{pmatrix}385&241\\241&891\end{pmatrix}\begin{bmatrix}6/17\\12/17\end{bmatrix}=-\begin{bmatrix}9/29\\21/29\end{bmatrix}$$

再由一维搜索 $\min\limits_{\lambda} f(\boldsymbol{X}^{(1)}+\lambda\boldsymbol{P}^{(1)})$，得

$$\lambda_1=\frac{29}{17}$$

从而

$$\boldsymbol{X}^{(2)}=\boldsymbol{X}^{(1)}+\lambda_1\boldsymbol{P}^{(1)}=\begin{bmatrix}26/17\\38/17\end{bmatrix}+\frac{29}{17}\begin{bmatrix}-9/29\\-21/29\end{bmatrix}=\begin{pmatrix}1\\1\end{pmatrix}$$

$$\nabla f(\boldsymbol{X}^{(2)})=(0,0)^{\mathrm{T}}$$

可知 $\boldsymbol{X}^{(2)}=(1,1)^{\mathrm{T}}$ 为极小点。

在以上讨论中，我们取第一个尺度矩阵 $\overline{\boldsymbol{H}}^{(0)}$ 为对称正定阵，以后的尺度矩阵由

式(7-81)逐步形成。可以证明，这样构成的尺度矩阵均为对称正定阵。由此可知其搜索方向 $\boldsymbol{P}^{(k)}=-\overline{\boldsymbol{H}}^{(k)}\nabla f(\boldsymbol{X}^{(k)})$为下降方向，这就可以保证每次迭代均能使目标函数值有所改善。

当把 DFP 变尺度法用于正定二次函数时，产生的搜索方向为共轭方向，因而也具有有限步收敛的性质。若将初始尺度矩阵也取为单位矩阵，对这种函数来说，DFP 法就与共轭梯度法一样了。

还要指出，可以采用不同的方法来构造尺度矩阵 $\overline{\boldsymbol{H}}^{(k)}$，从而就有不同的变尺度法。DFP 法属于拟牛顿法的一种。开始时取 $\overline{\boldsymbol{H}}^{(0)}=\boldsymbol{I}$，这相当于第一步采用最速下降法。以后的 $\overline{\boldsymbol{H}}^{(k)}$接近于 $\boldsymbol{H}(\boldsymbol{X}^{(k)})^{-1}$，当达到极小点时，从理论上讲，这时的尺度矩阵应等于该点处海赛阵的逆阵。

例 10 试用 DFP 法求

$$\min f(\boldsymbol{X})=4(x_1-5)^2+(x_2-6)^2$$

解

取

$$\overline{\boldsymbol{H}}^{(0)}=\begin{pmatrix}1&0\\0&1\end{pmatrix},\qquad \boldsymbol{X}^{(0)}=\begin{pmatrix}8\\9\end{pmatrix}$$

由于

$$\nabla f(\boldsymbol{X})=[8(x_1-5),2(x_2-6)]^{\mathrm{T}}$$

$$\nabla f(\boldsymbol{X}^{(0)})=(24,6)^{\mathrm{T}}$$

故

$$\begin{aligned}\boldsymbol{X}^{(1)}&=\boldsymbol{X}^{(0)}+\lambda_0\boldsymbol{P}^{(0)}=\boldsymbol{X}^{(0)}+\lambda_0[-\overline{\boldsymbol{H}}^{(0)}\nabla f(\boldsymbol{X}^{(0)})]\\&=\begin{pmatrix}8\\9\end{pmatrix}-\lambda_0\begin{pmatrix}1&0\\0&1\end{pmatrix}\begin{pmatrix}24\\6\end{pmatrix}=\begin{pmatrix}8\\9\end{pmatrix}-\lambda_0\begin{pmatrix}24\\6\end{pmatrix}\\&=\begin{pmatrix}8-24\lambda_0\\9-6\lambda_0\end{pmatrix}\end{aligned}$$

$$f(\boldsymbol{X}^{(1)})=4[(8-24\lambda_0)-5]^2+[(9-6\lambda_0)-6]^2$$

令

$$\frac{\mathrm{d}f(\boldsymbol{X}^{(1)})}{\mathrm{d}\lambda_0}=0$$

可得

$$\lambda_0=\frac{17}{130}$$

$$\boldsymbol{X}^{(1)}=[(8-24\lambda_0),(9-6\lambda_0)]^{\mathrm{T}}=(4.862,8.215)^{\mathrm{T}}$$

$$\Delta\boldsymbol{X}^{(0)}=\boldsymbol{X}^{(1)}-\boldsymbol{X}^{(0)}=(-3.138,-0.785)^{\mathrm{T}}$$

$$f(\boldsymbol{X}^{(1)})=4.985$$

$$\nabla f(\boldsymbol{X}^{(1)})=(-1.108,4.431)^{\mathrm{T}}$$

$$\Delta \boldsymbol{G}^{(0)}=\nabla f(\boldsymbol{X}^{(1)})-\nabla f(\boldsymbol{X}^{(0)})=(-25.108,-1.569)^{\mathrm{T}}$$

由此可得

$$\overline{\boldsymbol{H}}^{(1)}=\overline{\boldsymbol{H}}^{(0)}+\frac{\Delta \boldsymbol{X}^{(0)}(\Delta \boldsymbol{X}^{(0)})^{\mathrm{T}}}{(\Delta \boldsymbol{G}^{(0)})^{\mathrm{T}}\Delta \boldsymbol{X}^{(0)}}-\frac{\overline{\boldsymbol{H}}^{(0)}\Delta \boldsymbol{G}^{(0)}(\Delta \boldsymbol{G}^{(0)})^{\mathrm{T}}\overline{\boldsymbol{H}}^{(0)}}{(\Delta \boldsymbol{G}^{(0)})^{\mathrm{T}}\overline{\boldsymbol{H}}^{(0)}\Delta \boldsymbol{G}^{(0)}}$$

$$=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}+\frac{(-3.138,-0.785)^{\mathrm{T}}(-3.138,-0.785)}{(-25.108,-1.569)(-3.138,-0.785)^{\mathrm{T}}}-$$

$$\frac{\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}(-25.108,-1.569)^{\mathrm{T}}(-25.108,-1.569)\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}}{(-25.108,-1.569)\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}(-25.108,-1.569)^{\mathrm{T}}}$$

$$=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}+\begin{pmatrix}0.1231 & 0.0308\\ 0.0308 & 0.0077\end{pmatrix}-\begin{pmatrix}0.9961 & 0.0622\\ 0.0622 & 0.0039\end{pmatrix}$$

$$=\begin{pmatrix}0.1270 & -0.0315\\ -0.0315 & 1.0038\end{pmatrix}$$

故

$$\boldsymbol{X}^{(2)}=\boldsymbol{X}^{(1)}-\lambda_1\overline{\boldsymbol{H}}^{(1)}\nabla f(\boldsymbol{X}^{(1)})$$

$$=\begin{pmatrix}4.862\\ 8.215\end{pmatrix}-\lambda_1\begin{pmatrix}0.1270 & -0.0315\\ -0.0315 & 1.0038\end{pmatrix}\begin{pmatrix}-1.108\\ 4.431\end{pmatrix}$$

如上求最佳步长，可得

$$\lambda_1=0.4942$$

代入上式得

$$\boldsymbol{X}^{(2)}=(5,6)^{\mathrm{T}}$$

这就是极小点。

若将该问题的目标函数 $f(\boldsymbol{X})$ 表示成式(7-50)的形式，可知

$$\boldsymbol{A}=\begin{pmatrix}8 & 0\\ 0 & 2\end{pmatrix}$$

而

$$\boldsymbol{A}^{-1}=\begin{pmatrix}1/8 & 0\\ 0 & 1/2\end{pmatrix}$$

现计算出该问题的 $\overline{\boldsymbol{H}}^{(2)}$

$$\overline{\boldsymbol{H}}^{(2)}=\begin{pmatrix}1.25\times 10^{-1} & -8.882\times 10^{-16}\\ -8.882\times 10^{-16} & 5.00\times 10^{-1}\end{pmatrix}$$

可知二者实际相等。

在以上几节中，我们介绍了求解无约束极值问题的解析法，这些方法只是众多算法中的一部分。一般认为，从迭代次数上考虑，变尺度法所需迭代次数较少，共轭梯度法次之，

最速下降法(一阶梯度法)所需迭代次数最多。但从每次迭代所需的计算工作量来看，却正好相反，最速下降法最简单，变尺度法比它们都繁。

7.3.4 步长加速法

步长加速法亦称模式法或**模矢法**(Pattern Search)，由胡克(Hooke)和基夫斯(Jeeves)于1961年提出，是一种直接法。它易于编制计算机程序，且具有追寻谷线(脊线)加速移向最优点的性质。

1. 基本原理

假定欲求某实值函数 $f(\boldsymbol{X})$ 的极小点，为此，任选一**基点** $\boldsymbol{B}_1$(初始近似点)，算出此点的目标函数值。然后沿第 i 个坐标方向以某一步长 $\boldsymbol{\Delta}_i$ 进行探索，即在 $\boldsymbol{B}_1+\boldsymbol{\Delta}_i$ 和 $\boldsymbol{B}_1-\boldsymbol{\Delta}_i$ 这两点中寻求能使目标函数值下降的点，并把它作为**临时矢点**；再由此点出发沿另一坐标方向进行同样的探索，如能得到比以前更好的点，就以该点代替前面的点作为新的临时矢点。如此沿各个坐标方向轮流探查一遍，并选这一轮探索得到的最好的点(最后的临时矢点)为第二个基点 $\boldsymbol{B}_2$。由第一个基点 $\boldsymbol{B}_1$ 到第二个基点 $\boldsymbol{B}_2$ 构成了第一个**模矢**。对第一个基点来说，可以认为这个模矢的方向可能是使目标函数值得以改善的最有利的移动方向，沿这一方向前进，目标函数值下降"最快"(就 $\boldsymbol{B}_1$ 附近而言)。显然，这一方向近似于目标函数的负梯度方向(从而可知这一方法为近似最速下降法)。现假定在第二个基点 $\boldsymbol{B}_2$ 附近进行类似的探索，其结果可能和在 $\boldsymbol{B}_1$ 处的情形相同，故略去这步探索而把第一个模矢加长一倍(即所谓**加速**)。现设其端点 $\boldsymbol{T}_{20}$ 是第二个模矢的终点(下一步迭代的初始临时矢点)，这样 $\boldsymbol{B}_2\boldsymbol{T}_{20}$ 就构成了假定的第二个模矢。然后，在 $\boldsymbol{T}_{20}$ 附近进行如上类似的探索，得出新的最好的点——第三个基点 $\boldsymbol{B}_3$。据此修改假定的第二个模矢，使它的起点为 $\boldsymbol{B}_2$，终点为 $\boldsymbol{B}_3$。其后，再把第二个模矢延长一倍……如此继续进行**探索和加速**，即可得到越来越好的目标函数下降点(见图7-17)。

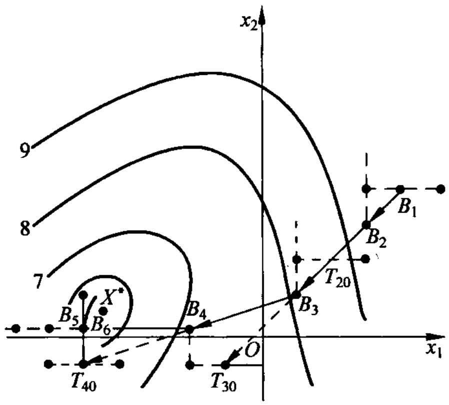


图 7-17

如果探索进行到某一步时得不出新的下降点，则应缩小步长以进行更精细的探索。当步长已缩小到某一精度要求，但仍得不到新的下降点时，即可将该点作为所求的近似极小点，就此停止迭代。

图7-17示出了一个在二维空间中用模矢法探求极小点 $\boldsymbol{X}^*$ 的例子。该例从初始基点 $\boldsymbol{B}_1$ 开始，用模矢法相继得出基点 $\boldsymbol{B}_2$、$\boldsymbol{B}_3$、$\boldsymbol{B}_4$ 和 $\boldsymbol{B}_5$。这时如不缩小步长，就得不出新的基点(图中的基点 $\boldsymbol{B}_6$ 和 $\boldsymbol{B}_5$ 是同一个点)。

2. 计算步骤

根据上述分析，现将用模矢法求解无约束极值问题式(7-40)的计算步骤总结如下。

(1) 任选初始近似点 $\boldsymbol{B}_1$，以它为初始基点进行探索。

(2) 为每一独立变量 $x_i(i=1,2,\cdots,n)$选定步长

$$\boldsymbol{\Delta}_i=\begin{bmatrix}0\\ \vdots\\ 0\\ \delta_i\\ 0\\ \vdots\\ 0\end{bmatrix}\leftarrow \text{第 } i \text{ 个分量} \tag{7-82}$$

上式中$\boldsymbol{\Delta}_i$ 为第 i 个分量是 δ_i，而其他所有分量均为零的向量。

(3) 算出初始基点 $\boldsymbol{B}_1$ 的目标函数值 $f(\boldsymbol{B}_1)$，考虑点 $\boldsymbol{B}_1+\boldsymbol{\Delta}_1$，若 $f(\boldsymbol{B}_1+\boldsymbol{\Delta}_1)<f(\boldsymbol{B}_1)$，就以 $\boldsymbol{B}_1+\boldsymbol{\Delta}_1$ 为临时矢点，并记为 $\boldsymbol{T}_{11}$。这里的第一个下标表示现在是在建立第一个模矢，第二个下标表示变量 x_1 已被摄动。若 $\boldsymbol{B}_1+\boldsymbol{\Delta}_1$ 不比 $\boldsymbol{B}_1$ 点好，就试验$\boldsymbol{B}_1-\boldsymbol{\Delta}_1$，如果它比 $\boldsymbol{B}_1$ 点好，就以它为临时矢点，否则，以 $\boldsymbol{B}_1$ 为临时矢点。即

$$\boldsymbol{T}_{11}=\begin{cases}\boldsymbol{B}_1+\boldsymbol{\Delta}_1, \text{若 } f(\boldsymbol{B}_1+\boldsymbol{\Delta}_1)<f(\boldsymbol{B}_1)\\ \boldsymbol{B}_1-\boldsymbol{\Delta}_1, \text{若 } f(\boldsymbol{B}_1-\boldsymbol{\Delta}_1)<f(\boldsymbol{B}_1)\leqslant f(\boldsymbol{B}_1+\boldsymbol{\Delta}_1)\\ \boldsymbol{B}_1, \qquad \text{若 } f(\boldsymbol{B}_1)\leqslant \min[f(\boldsymbol{B}_1+\boldsymbol{\Delta}_1),f(\boldsymbol{B}_1-\boldsymbol{\Delta}_1)]\end{cases} \tag{7-83}$$

对于下一个独立变量 x_2 进行类似的摄动，这时，用临时矢点 $\boldsymbol{T}_{11}$代替原来的基点 $\boldsymbol{B}_1$。一般地

$$\boldsymbol{T}_{1,j+1}=\begin{cases}\boldsymbol{T}_{1j}+\boldsymbol{\Delta}_{j+1}, \text{若 } f(\boldsymbol{T}_{1j}+\boldsymbol{\Delta}_{j+1})<f(\boldsymbol{T}_{1j})\\ \boldsymbol{T}_{1j}-\boldsymbol{\Delta}_{j+1}, \text{若 } f(\boldsymbol{T}_{1j}-\boldsymbol{\Delta}_{j+1})<f(\boldsymbol{T}_{1j})\leqslant f(\boldsymbol{T}_{1j}+\boldsymbol{\Delta}_{j+1})\\ \boldsymbol{T}_{1j}, \qquad \text{若 } f(\boldsymbol{T}_{1j})\leqslant \min[f(\boldsymbol{T}_{1j}+\boldsymbol{\Delta}_{j+1}),f(\boldsymbol{T}_{1j}-\boldsymbol{\Delta}_{j+1})]\end{cases} \tag{7-84}$$

上式中，$0\leqslant j\leqslant n-1$，$\boldsymbol{T}_{10}=\boldsymbol{B}_1$。

n 个变量都摄动之后，得临时矢点 $\boldsymbol{T}_{1n}$，并令

$$\boldsymbol{T}_{1n}=\boldsymbol{B}_2 \tag{7-85}$$

原来的基点 $\boldsymbol{B}_1$ 和新基点 $\boldsymbol{B}_2$ 确定了第一个模矢。

(4) 将第一个模矢延长一倍，得第二个模矢的初始临时矢点 $\boldsymbol{T}_{20}$

$$\boldsymbol{T}_{20}=\boldsymbol{B}_1+2(\boldsymbol{B}_2-\boldsymbol{B}_1)=2\boldsymbol{B}_2-\boldsymbol{B}_1 \tag{7-86}$$

(5) 在 $\boldsymbol{T}_{20}$附近进行和上面类似的探索，建立临时矢点 $\boldsymbol{T}_{21},\boldsymbol{T}_{22},\cdots,\boldsymbol{T}_{2n}$，以 $\boldsymbol{T}_{2n}$为第三个基点 $\boldsymbol{B}_3$。这样，$\boldsymbol{B}_2$、$\boldsymbol{B}_3$ 就确立了第二个模矢。第三个模矢的初始临时矢点为

$$\boldsymbol{T}_{30}=\boldsymbol{B}_2+2(\boldsymbol{B}_3-\boldsymbol{B}_2)=2\boldsymbol{B}_3-\boldsymbol{B}_2$$

注意，在进行如上探索时，若在一个方向上重复见效，就会使模矢增长，这一点可由下式看出

$$\boldsymbol{B}_3-\boldsymbol{B}_2=2(\boldsymbol{T}_{20}-\boldsymbol{B}_2)=2(\boldsymbol{B}_2-\boldsymbol{B}_1) \tag{7-87}$$

(6) 继续上述过程。对于第 i 个模矢,如果

$$f(\boldsymbol{T}_{i0})<f(\boldsymbol{B}_i)$$

但沿各坐标方向的所有摄动均得不出比 $\boldsymbol{T}_{i0}$ 更好的点,则以 $\boldsymbol{T}_{i0}$ 为 $\boldsymbol{B}_{i+1}$,而且不把这个模矢延长。

若 $f(\boldsymbol{T}_{i0})\geqslant f(\boldsymbol{B}_i)$,且由 $\boldsymbol{T}_{i0}$ 产生不出比 $\boldsymbol{B}_i$ 更好的点,则应退回到 $\boldsymbol{B}_i$,并在 $\boldsymbol{B}_i$ 附近进行探索。如能得出新的下降点,即可引出新的模矢;否则,将步长缩小,以进行更精细的探查。当步长缩小到要求的精度时,即可停止迭代。

对于比较复杂的目标函数,为了防止把局部极值误认为全局最优值,应分区域进行探查,并从各区域搜索得到的局部极值和极值点中选取最优者;或者从任意选取的不同点开始,至少引入两个独立的搜索,如果它们都收敛于同一个点,则这个点作为最优点的把握就大大增加了。

CHAPTER 8 第8章*

约束极值问题

实际工作中遇到的大多数极值问题，其变量的取值多受到一定限制，这种限制由约束条件来体现。带有约束条件的极值问题称为约束极值问题，也叫规划问题。非线性规划的一般形式为

$$\begin{cases}\min f(\boldsymbol{X}) \\ \quad h_i(\boldsymbol{X})=0, \quad i=1,2,\cdots,m \\ \quad g_j(\boldsymbol{X})\geqslant 0, \quad j=1,2,\cdots,l\end{cases} \tag{8-1}$$

或

$$\begin{cases}\min f(\boldsymbol{X}) \\ \quad g_j(\boldsymbol{X})\geqslant 0, \quad j=1,2,\cdots,l\end{cases} \tag{8-2}$$

问题式(8-2)也常写成

$$\begin{cases}\min f(\boldsymbol{X}),\boldsymbol{X}\in R\subset E^n \\ \quad R=\{\boldsymbol{X}\,|\,g_j(\boldsymbol{X})\geqslant 0, \quad j=1,2,\cdots,l\}\end{cases} \tag{8-3}$$

求解约束极值问题要比求解无约束极值问题困难得多。对有约束的极小化问题来说，除了要使目标函数在每次迭代有所下降之外，还要时刻注意解的可行性问题（某些算法的中间步骤除外），这就给寻优工作带来了很大困难。为了实际求解和（或）简化其优化工作，可采用以下方法：将约束问题化为无约束问题；将非线性规划问题化为线性规划问题，以及能将复杂问题变换为较简单问题的其他方法。

8.1 最优性条件

8.1.1 起作用约束和可行下降方向的概念

现考虑上述一般非线性规划，假定 $f(\boldsymbol{X})$、$h_i(\boldsymbol{X})$ 和 $g_j(\boldsymbol{X})$ $(i=1,2,\cdots,m;\ j=1,2,\cdots,l)$ 具有一阶连续偏导数。

设 $\boldsymbol{X}^{(0)}$ 是非线性规划的一个可行解，它当然满足所有约束。现考虑某一不等式约束条件 $g_j(\boldsymbol{X})\geqslant 0$，$\boldsymbol{X}^{(0)}$ 满足它有两种可能：其一为 $g_j(\boldsymbol{X}^{(0)})>0$，这时，点 $\boldsymbol{X}^{(0)}$ 不是处于由这一约束条件形成的可行域边界上，因而这一约束对 $\boldsymbol{X}^{(0)}$ 点的微小摄动不起限制作用，从而称这个约束条件是 $\boldsymbol{X}^{(0)}$ 点的**不起作用约束**（或无效约束）；其二是 $g_j(\boldsymbol{X}^{(0)})=0$，这时 $\boldsymbol{X}^{(0)}$

点处于该约束条件形成的可行域边界上，它对 $X^{(0)}$ 的摄动起到了某种限制作用，故称这个约束是 $X^{(0)}$ 点的**起作用约束**（有效约束）。

显而易见，等式约束对所有可行点来说都是起作用约束。

假定 $X^{(0)}$ 是非线性规划式(8-3)的一个可行点，现考虑此点的某一方向 D，若存在实数 $\lambda_0>0$，使对任意 $\lambda\in[0,\lambda_0]$均有

$$X^{(0)}+\lambda D\in R \tag{8-4}$$

就称方向 D 是 $X^{(0)}$ 点的一个**可行方向**。

若 D 是可行点 $X^{(0)}$ 处的任一可行方向（参看图 8-1），则对该点的所有起作用约束。

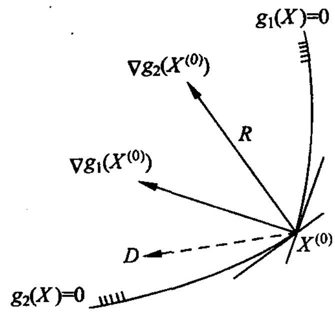


图 8-1

$$g_j(X)\geqslant 0$$

均有

$$\nabla g_j(X^{(0)})^{\mathrm{T}}D\geqslant 0,\quad j\in J \tag{8-5}$$

其中 J 为这个点所有起作用约束下标的集合。

另一方面，由泰勒公式

$$g_j(X^{(0)}+\lambda D)=g_j(X^{(0)})+\lambda\,\nabla g_j(X^{(0)})^{\mathrm{T}}D+o(\lambda)$$

可知对所有起作用约束，当 $\lambda>0$ 足够小时，只要

$$\nabla g_j(X^{(0)})^{\mathrm{T}}D>0,\quad j\in J \tag{8-6}$$

就有

$$g_j(X^{(0)}+\lambda D)\geqslant 0,\quad j\in J$$

此外，对 $X^{(0)}$ 点的不起作用约束，由约束函数的连续性，当 $\lambda>0$ 足够小时亦有上式成立。从而，只要方向 D 满足式(8-6)，即可保证它是 $X^{(0)}$ 点的可行方向。

考虑非线性规划的某一可行点 $X^{(0)}$，对该点的任一方向 D 来说，若存在实数 $\lambda_0'>0$，使对任意 $\lambda\in[0,\lambda_0']$均有

$$f(X^{(0)}+\lambda D)<f(X^{(0)})$$

就称方向 D 为 $X^{(0)}$ 点的一个**下降方向**。

将目标函数 $f(X)$在点 $X^{(0)}$ 处作一阶泰勒展开，可知满足条件

$$\nabla f(X^{(0)})^{\mathrm{T}}D<0 \tag{8-7}$$

的方向 D 必为 $X^{(0)}$ 点的下降方向。

如果方向 D 既是 $X^{(0)}$ 点的可行方向，又是这个点的下降方向，就称它是该点的**可行下降方向**。假如 $X^{(0)}$ 点不是极小点，继续寻优时的搜索方向就应从该点的可行下降方向中去找。显然，若某点存在可行下降方向，它就不会是极小点。另外，若某点为极小点，则在该点不存在可行下降方向。

定理 1 设 X^* 是非线性规划式(8-3)的一个局部极小点，目标函数 $f(X)$在 X^* 处可微，而且

$g_j(X)$在 X^* 处可微，当 $j\in J$

$g_j(X)$在 X^* 处连续，当 $j\overline{\in}J$

则在 X^* 点不存在可行下降方向，从而不存在向量 D 同时满足：

$$\begin{cases}\nabla f(X^*)^{\mathrm{T}}D<0 \\ \nabla g_j(X^*)^{\mathrm{T}}D>0, \quad j\in J\end{cases} \tag{8-8}$$

这个定理显然是成立的。事实上，若存在满足式(8-8)的方向 D，则沿该方向搜索可找到更好的可行点，从而与 X^* 为极小点的假设矛盾。

式(8-8)的几何意义十分明显。满足该条件的方向 D，与点 X^* 处目标函数负梯度方向的夹角为锐角，与点 X^* 处起作用约束梯度方向的夹角也为锐角。

8.1.2 库恩-塔克条件

假定 X^* 是非线性规划式(8-3)的极小点，该点可能位于可行域的内部，也可能处于可行域的边界上。若为前者，这事实上是个无约束问题，X^* 必满足条件 $\nabla f(X^*)=0$；若为后者，情况就复杂得多了，现在我们来讨论后一种情形。

不失一般性，设 X^* 位于第一个约束条件形成的可行域边界上，即第一个约束条件是 X^* 点的起作用约束($g_1(X^*)=0$)。若 X^* 是极小点，则 $\nabla g_1(X^*)$ 必与 $-\nabla f(X^*)$ 在一条直线上且方向相反(我们在这里假定向量 $\nabla g_1(X^*)$ 和 $\nabla f(X^*)$ 皆不为零)，否则，在该点就一定存在可行下降方向(图 8-2 中的 X^* 点为极小点；X 点不满足上述要求，它不是极小点，角度 β 表示了该点可行下降方向的范围)。上面的论述说明，在上述条件下，存在实数 $\gamma_1\geqslant 0$，使

$$\nabla f(X^*)-\gamma_1\nabla g_1(X^*)=0$$

若 X^* 点有两个起作用约束，例如说有 $g_1(X^*)=0$ 和 $g_2(X^*)=0$。在这种情况下，$\nabla f(X^*)$ 必处于 $\nabla g_1(X^*)$ 和 $\nabla g_2(X^*)$ 的夹角之内。如若不然，在 X^* 点必有可行下降方向，它就不会是极小点(图 8-3)。由此可见，如果 X^* 是极小点，而且 X^* 点的起作用约束条件的梯度 $\nabla g_1(X^*)$ 和 $\nabla g_2(X^*)$ 线性无关，则可将 $\nabla f(X^*)$ 表示成 $\nabla g_1(X^*)$ 和 $\nabla g_2(X^*)$ 的非负线性组合。也就是说，在这种情况下存在实数 $\gamma_1\geqslant 0$ 和 $\gamma_2\geqslant 0$，使

$$\nabla f(X^*)-\gamma_1\nabla g_1(X^*)-\gamma_2\nabla g_2(X^*)=0$$

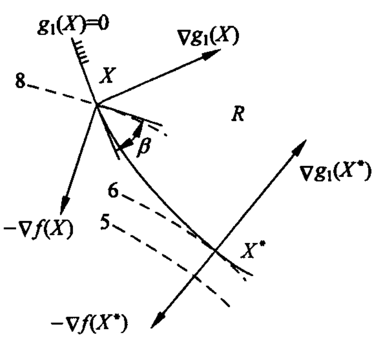


图 8-2

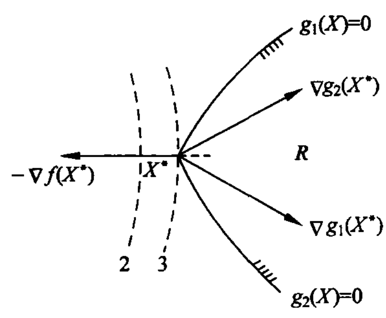


图 8-3

如上类推,可以得到

$$\nabla f(\boldsymbol{X}^*)-\sum_{j\in J}\gamma_j\,\nabla g_j(\boldsymbol{X}^*)=0 \tag{8-9}$$

为了把不起作用约束也包括进式(8-9)中,增加条件

$$\begin{cases}\gamma_j\,g_j(\boldsymbol{X}^*)=0\\ \gamma_j\geqslant 0\end{cases}$$

当 $g_j(\boldsymbol{X}^*)=0$ 时,γ_j 可不为零;当 $g_j(\boldsymbol{X}^*)\neq 0$ 时,必有 $\gamma_j=0$。如此即可得到著名的**库恩-塔克(Kuhn-Tucker,简写为 K-T)条件**。

库恩-塔克条件是非线性规划领域中最重要的理论成果之一,是确定某点为最优点的必要条件。只要是最优点(而且该点起作用约束的梯度线性无关,满足这种要求的点称为正则点),就必须满足这个条件。但一般说它并不是充分条件,因而满足这个条件的点不一定就是最优点(对于凸规划,它既是最优点存在的必要条件,同时也是充分条件)。

现可将**库恩-塔克条件**叙述如下:

设 $\boldsymbol{X}^*$ 是非线性规划式(8-3)的极小点,而且在 $\boldsymbol{X}^*$ 点的各起作用约束的梯度线性无关,则存在向量 $\boldsymbol{\Gamma}^*=(\gamma_1^*,\gamma_2^*,\cdots,\gamma_l^*)^{\mathrm{T}}$,使下述条件成立:

$$\begin{cases}\nabla f(\boldsymbol{X}^*)-\sum\limits_{j=1}^{l}\gamma_j^*\ \nabla g_j(\boldsymbol{X}^*)=0\\ \gamma_j^*\,g_j(\boldsymbol{X}^*)=0,\quad j=1,2,\cdots,l\\ \gamma_j^*\geqslant 0,\qquad j=1,2,\cdots,l\end{cases} \tag{8-10}$$

条件式(8-10)常简称为 K-T 条件。满足这个条件的点(它当然也满足非线性规划的所有约束条件)称为**库恩-塔克点**(或 K-T 点)。

为了得出非线性规划式(8-1)的库恩-塔克条件,我们用

$$\begin{cases}h_i(\boldsymbol{X})\geqslant 0\\ -h_i(\boldsymbol{X})\geqslant 0\end{cases}$$

代替约束条件 $h_i(\boldsymbol{X})=0$,这样即可使用条件式(8-10),从而得到这时的**库恩-塔克条件**如下:

设 $\boldsymbol{X}^*$ 是非线性规划式(8-1)的极小点,而且 $\boldsymbol{X}^*$ 点的所有起作用约束的梯度 $\nabla h_i(\boldsymbol{X}^*)(i=1,2,\cdots,m)$ 和 $\nabla g_j(\boldsymbol{X}^*)(j\in J)$ **线性无关**,则存在向量

$$\boldsymbol{\Lambda}^*=(\lambda_1^*,\lambda_2^*,\cdots,\lambda_m^*)^{\mathrm{T}} \text{ 和 } \boldsymbol{\Gamma}^*=(\gamma_1^*,\gamma_2^*,\cdots,\gamma_l^*)^{\mathrm{T}}$$

使下述条件成立

$$\begin{cases}\nabla f(\boldsymbol{X}^*)-\sum\limits_{i=1}^{m}\lambda_i^*\ \nabla h_i(\boldsymbol{X}^*)-\sum\limits_{j=1}^{l}\gamma_j^*\ \nabla g_j(\boldsymbol{X}^*)=0\\ \gamma_j^*\,g_j(\boldsymbol{X}^*)=0,\quad j=1,2,\cdots,l\\ \gamma_j^*\geqslant 0,\qquad j=1,2,\cdots,l\end{cases} \tag{8-11}$$

当然,满足条件式(8-11)的点也称为库恩-塔克点。

在条件式(8-10)和式(8-11)中,$\lambda_1^*,\lambda_2^*,\cdots,\lambda_m^*$ 以及 $\gamma_1^*,\gamma_2^*,\cdots,\gamma_l^*$ 称为**广义拉格朗日(Lagrange)乘子**。

例 1　用库恩-塔克条件解非线性规划

$$\begin{cases}\min f(x)=(x-3)^2\\ \quad 0\leqslant x\leqslant 5\end{cases}$$

解　先将该非线性规划问题写成以下形式

$$\begin{cases}\min f(x)=(x-3)^2\\ \quad g_1(x)=x\geqslant 0\\ \quad g_2(x)=5-x\geqslant 0\end{cases}$$

写出其目标函数和约束函数的梯度

$$\nabla f(x)=2(x-3),$$
$$\nabla g_1(x)=1,\nabla g_2(x)=-1$$

对第一个和第二个约束条件分别引入广义拉格朗日乘子 γ_1^* 和 γ_2^*,设 K-T 点为 x^*,则可写出该问题的 K-T 条件如下

$$\begin{cases}2(x^*-3)-\gamma_1^*+\gamma_2^*=0\\ \gamma_1^*x^*=0\\ \gamma_2^*(5-x^*)=0\\ \gamma_1^*,\gamma_2^*\geqslant 0\end{cases}$$

为解上述方程组,考虑以下几种情形:

(1) 令 $\gamma_1^*\neq 0,\gamma_2^*\neq 0$,无解。

(2) 令 $\gamma_1^*\neq 0,\gamma_2^*=0$,解之,得 $x^*=0,\gamma_1^*=-6$,不是 K-T 点。

(3) 令 $\gamma_1^*=0,\gamma_2^*\neq 0$,解之,得 $x^*=5,\gamma_2^*=-4$,不是 K-T 点。

(4) 令 $\gamma_1^*=\gamma_2^*=0$,解之,得 $x^*=3$,此为 K-T 点,其目标函数值 $f(x^*)=0$。

由于该非线性规划问题为凸规划,故 $x^*=3$ 就是其全局极小点。该点是可行域的内点,它也可直接由梯度等于零的条件求出。

8.2　二次规划

若某非线性规划的目标函数为自变量 X 的二次函数,约束条件又全是线性的,就称这种规划为**二次规划**。二次规划是非线性规划中比较简单的一类,它较容易求解。由于很多方面的问题都可以抽象成二次规划的模型,而且它和线性规划又有直接联系,因而此处专门提出来简要做一说明。

二次规划的数学模型可表述如下

$$\begin{cases} \min f(\boldsymbol{X}) = \sum_{j=1}^{n} c_j x_j + \frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n} c_{jk} x_j x_k & (8\text{-}12) \\ c_{jk} = c_{kj}, \qquad k = 1,2,\cdots,n & \\ \sum_{j=1}^{n} a_{ij} x_j + b_i \geqslant 0, \quad i = 1,2,\cdots,m & (8\text{-}13) \\ x_j \geqslant 0, \qquad j = 1,2,\cdots,n & (8\text{-}14) \end{cases}$$

式(8-12)右端的第二项为二次型。如果该二次型正定(或半正定),则目标函数为严格凸函数(或凸函数);此外,二次规划的可行域为凸集,因而,上述规划属于凸规划(在极大化问题中,如果上述二次型为负定或半负定,则也属于凸规划)。在本章中已经指出,凸规划的局部极值即为其全局极值。对于这种问题来说,库恩-塔克条件不但是极值点存在的必要条件,而且也是充分条件。

将库恩-塔克条件式(8-10)中的第一个条件应用于二次规划式(8-12)~式(8-14),并用 y 代替库恩-塔克条件中的 γ,即可得到

$$-\sum_{k=1}^{n} c_{jk} x_k + \sum_{i=1}^{m} a_{ij} y_{n+i} + y_j = c_j, \qquad j = 1,2,\cdots,n \tag{8-15}$$

在式(8-13)中引入松弛变量 x_{n+i},式(8-13)即变为(假定 $b_i \geqslant 0$)

$$\sum_{j=1}^{n} a_{ij} x_j - x_{n+i} + b_i = 0, \qquad i = 1,2,\cdots,m \tag{8-16}$$

再将库恩-塔克条件中的第二个条件应用于上述二次规划,并考虑到式(8-16),这就得到

$$x_j y_j = 0, \quad j = 1,2,\cdots,n+m \tag{8-17}$$

此外还有

$$x_j \geqslant 0, \quad y_j \geqslant 0, \quad j = 1,2,\cdots,n+m \tag{8-18}$$

联立求解式(8-15)和式(8-16),如果得到的解也满足式(8-17)和式(8-18),则这样的解就是原二次规划问题的解。但是,在式(8-15)中,c_j 可能为正,也可能为负。为了便于求解,先引入人工变量 z_j($z_j \geqslant 0$,其前面的符号可取正或负,以便得出可行解),这样式(8-15)就变成了

$$\sum_{i=1}^{m} a_{ij} y_{n+i} + y_j - \sum_{k=1}^{n} c_{jk} x_k + \operatorname{sgn}(c_j) z_j = c_j, \quad j = 1,2,\cdots,n \tag{8-19}$$

其中 $\operatorname{sgn}(c_j)$ 为符号函数,当 $c_j \geqslant 0$ 时,$\operatorname{sgn}(c_j) = 1$;当 $c_j < 0$ 时,$\operatorname{sgn}(c_j) = -1$。这样一来,可立刻得到初始基本可行解如下

$$\begin{cases} z_j = \operatorname{sgn}(c_j) c_j, & j = 1,2,\cdots,n \\ x_{n+i} = b_i, & i = 1,2,\cdots,m \\ x_j = 0, & j = 1,2,\cdots,n \\ y_j = 0, & j = 1,2,\cdots,n+m \end{cases}$$

但是,只有当 $z_j = 0$ 时才能得到原来问题的解,故必须对上述问题进行修正,从而得

到如下线性规划问题

$$\begin{cases}\min \varphi(Z) = \sum_{j=1}^{n} z_j \\ \sum_{i=1}^{m} a_{ij} y_{n+i} + y_j - \sum_{k=1}^{n} c_{jk} x_k + \operatorname{sgn}(c_j) z_j = c_j\ , \quad j=1,2,\cdots,n \\ \sum_{j=1}^{n} a_{ij} x_j - x_{n+i} + b_i = 0\ , \quad i=1,2,\cdots,m \\ x_j \geqslant 0 \quad j=1,2,\cdots,n+m \\ y_j \geqslant 0 \quad j=1,2,\cdots,n+m \\ z_j \geqslant 0 \quad j=1,2,\cdots,n \end{cases} \tag{8-20}$$

该线性规划尚应满足式(8-17)。这相当于说，不能使 x_j 和 y_j(对每一个 j)同时为基变量。解线性规划式(8-20)，若得到最优解

$$(x_1^*, x_2^*, \cdots, x_{n+m}^*, y_1^*, y_2^*, \cdots, y_{n+m}^*, z_1=0, z_2=0, \cdots, z_n=0)$$

则$(x_1^*, x_2^*, \cdots, x_n^*)$就是原二次规划问题的最优解。

例 2 求解二次规划

$$\begin{cases}\max f(\boldsymbol{X}) = 8x_1 + 10x_2 - x_1^2 - x_2^2 \\ 3x_1 + 2x_2 \leqslant 6 \\ x_1, x_2 \geqslant 0\end{cases}$$

解 将上述二次规划改写为

$$\begin{cases}\min \bar{f}(\boldsymbol{X}) = x_1^2 + x_2^2 - 8x_1 - 10x_2 = \dfrac{1}{2}(2x_1^2 + 2x_2^2) - 8x_1 - 10x_2 \\ 6 - 3x_1 - 2x_2 \geqslant 0 \\ x_1 \geqslant 0, \quad x_2 \geqslant 0\end{cases}$$

可知目标函数为严格凸函数。此外

$$c_1 = -8, \quad c_2 = -10, \quad c_{11} = 2, \quad c_{22} = 2,$$
$$c_{12} = c_{21} = 0, \quad b_1 = 6, \quad a_{11} = -3, \quad a_{12} = -2$$

由于 c_1 和 c_2 小于零，故引入的人工变量 z_1 和 z_2 前面取负号，这样就得到线性规划问题如下

$$\begin{cases}\min \varphi(Z) = z_1 + z_2 \\ -3y_3 + y_1 - 2x_1 - z_1 = -8 \\ -2y_3 + y_2 - 2x_2 - z_2 = -10 \\ -3x_1 - 2x_2 - x_3 + 6 = 0 \\ x_1, x_2, x_3, y_1, y_2, y_3, z_1, z_2 \geqslant 0\end{cases}$$

或

$$\begin{cases}\min \varphi(Z)=z_1+z_2 \\ \quad 2x_1+3y_3-y_1+z_1=8 \\ \quad 2x_2+2y_3-y_2+z_2=10 \\ \quad 3x_1+2x_2+x_3=6 \\ \quad x_1,x_2,x_3,y_1,y_2,y_3,z_1,z_2\geqslant 0\end{cases}$$

此外尚应满足

$$x_jy_j=0, \qquad j=1,2,3$$

用线性规划的单纯形法解之(注意,在转换过程中应满足条件 $x_jy_j=0$),得该线性规划问题的解如下

$$x_1=4/13, \quad x_2=33/13, \quad x_3=0, \quad y_1=0,$$
$$y_2=0, \quad y_3=32/13, \quad z_1=0, \quad z_2=0$$

由此得到原二次规划问题的解为

$$x_1^*=4/13,\ x_2^*=33/13,\ f(X^*)=21.3$$

读者可以验证

$$x_1^*=4/13,\ x_2^*=33/13,\ \gamma_1^*=0,\ \gamma_2^*=0$$

以及 $\gamma_3^*=32/13$ 满足库恩-塔克条件。

8.3 可行方向法

现考虑非线性规划式(8-3),设 $\boldsymbol{X}^{(k)}$ 是它的一个可行解,但不是要求的极小点。为了求它的极小点或近似极小点,根据以前所说,应在 $\boldsymbol{X}^{(k)}$ 点的可行下降方向中选取某一方向 $\boldsymbol{D}^{(k)}$,并确定步长 λ_k,使

$$\begin{cases}\boldsymbol{X}^{(k+1)}=\boldsymbol{X}^{(k)}+\lambda_k\boldsymbol{D}^{(k)}\in R \\ f(\boldsymbol{X}^{(k+1)})<f(\boldsymbol{X}^{(k)})\end{cases} \tag{8-21}$$

若满足精度要求,迭代停止,$\boldsymbol{X}^{(k+1)}$ 就是所要的点。否则,从 $\boldsymbol{X}^{(k+1)}$ 出发继续进行迭代,直到满足要求为止。上述这种方法称为可行方向法,它具有下述特点:①迭代过程中所采用的搜索方向为可行方向;②所产生的迭代点列$\{\boldsymbol{X}^{(k)}\}$始终在可行域内;③目标函数值单调下降。由此可见,很多方法都可以归入可行方向法一类。但我们通常所说的可行方向法,一般指的是 Zoutendijk 在 1960 年提出的算法及其变形,下面就来说明 Zoutendijk 的可行方向法。

设 $\boldsymbol{X}^{(k)}$ 点的起作用约束集非空,为求 $\boldsymbol{X}^{(k)}$ 点的可行下降方向,可由下述不等式组确定向量 $\boldsymbol{D}$

$$\begin{cases}\nabla f(\boldsymbol{X}^{(k)})^{\mathrm{T}}\boldsymbol{D}<0 \\ \nabla g_j(\boldsymbol{X}^{(k)})^{\mathrm{T}}\boldsymbol{D}>0, \quad j\in J\end{cases} \tag{8-22}$$

这等价于由下面的不等式组求向量 $\boldsymbol{D}$ 和实数 η

$$\begin{cases}\nabla f(\boldsymbol{X}^{(k)})^{\mathrm{T}}\boldsymbol{D}\leqslant\eta \\ -\nabla g_j(\boldsymbol{X}^{(k)})^{\mathrm{T}}\boldsymbol{D}\leqslant\eta, \quad j\in J \\ \eta<0\end{cases} \tag{8-23}$$

现使 $\nabla f(\boldsymbol{X}^{(k)})^{\mathrm{T}}\boldsymbol{D}$ 和 $-\nabla g_j(\boldsymbol{X}^{(k)})^{\mathrm{T}}\boldsymbol{D}$（对所有 $j\in J$）的最大值极小化（必须同时限制向量 $\boldsymbol{D}$ 的模），即可将上述选取搜索方向的工作，转换为求解下述线性规划问题

$$\begin{cases}\min\ \eta \\ \quad\nabla f(\boldsymbol{X}^{(k)})^{\mathrm{T}}\boldsymbol{D}\leqslant\eta, \\ \quad-\nabla g_j(\boldsymbol{X}^{(k)})^{\mathrm{T}}\boldsymbol{D}\leqslant\eta, \quad j\in J(\boldsymbol{X}^{(k)}) \\ \quad-1\leqslant d_i\leqslant 1, \qquad i=1,2,\cdots,n\end{cases} \tag{8-24}$$

式中 $d_i(i=1,2,\cdots,n)$ 为向量 $\boldsymbol{D}$ 的分量。在式(8-24)中加入最后一个限制条件，为的是使该线性规划有**有限最优解**；由于我们的目的在于寻找搜索方向 $\boldsymbol{D}$，只需知道 $\boldsymbol{D}$ 的各分量的相对大小即可。

将线性规划式(8-24)的最优解记为 $(\boldsymbol{D}^{(k)},\eta_k)$，如果求出的 $\eta_k=0$，说明在 $\boldsymbol{X}^{(k)}$ 点不存在可行下降方向，在 $\nabla g_j(\boldsymbol{X}^{(k)})$（此处 $j\in J(\boldsymbol{X}^{(k)})$）线性无关的条件下，$\boldsymbol{X}^{(k)}$ 为一 K-T 点。若解出的 $\eta_k<0$，则得到可行下降方向 $\boldsymbol{D}^{(k)}$，这就是我们所要的搜索方向。

上述可行方向法的迭代步骤如下：

(1) 确定允许误差 $\varepsilon_1>0$ 和 $\varepsilon_2>0$，选初始近似点 $\boldsymbol{X}^{(0)}\in R$，并令 $k:=0$。

(2) 确定起作用约束指标集

$$J(\boldsymbol{X}^{(k)})=\{j\mid g_j(\boldsymbol{X}^{(k)})=0,1\leqslant j\leqslant l\}$$

① 若 $J(\boldsymbol{X}^{(k)})=\varnothing$（$\varnothing$ 为空集），而且 $\|\nabla f(\boldsymbol{X}^{(k)})\|^2\leqslant\varepsilon_1$，停止迭代，得点 $\boldsymbol{X}^{(k)}$。

② 若 $J(\boldsymbol{X}^{(k)})=\varnothing$，但 $\|\nabla f(\boldsymbol{X}^{(k)})\|^2>\varepsilon_1$，则取搜索方向 $\boldsymbol{D}^{(k)}=-\nabla f(\boldsymbol{X}^{(k)})$，然后转向第(5)步。

③ 若 $J(\boldsymbol{X}^{(k)})\neq\varnothing$，转下一步。

(3) 求解线性规划

$$\begin{cases}\min\ \eta \\ \quad\nabla f(\boldsymbol{X}^{(k)})^{\mathrm{T}}\boldsymbol{D}\leqslant\eta \\ \quad-\nabla g_j(\boldsymbol{X}^{(k)})^{\mathrm{T}}\boldsymbol{D}\leqslant\eta, \quad j\in J(\boldsymbol{X}^{(k)}) \\ \quad-1\leqslant d_i\leqslant 1, \qquad i=1,2,\cdots,n\end{cases}$$

设它的最优解是 $(\boldsymbol{D}^{(k)},\eta_k)$。

(4) 检验是否满足

$$|\eta_k|\leqslant\varepsilon_2$$

若满足则停止迭代，得到点 $\boldsymbol{X}^{(k)}$；否则，以 $\boldsymbol{D}^{(k)}$ 为搜索方向，并转下一步。

(5) 解下述一维极值问题

$$\lambda_k:\ \min_{0\leqslant\lambda\leqslant\bar{\lambda}} f(\boldsymbol{X}^{(k)}+\lambda\boldsymbol{D}^{(k)})$$

此处

$$\bar{\lambda}=\max\{\lambda\,|\,g_j(\boldsymbol{X}^{(k)}+\lambda\boldsymbol{D}^{(k)})\geqslant 0,\quad j=1,2,\cdots,l\}$$

(6) 令

$$\boldsymbol{X}^{(k+1)}=\boldsymbol{X}^{(k)}+\lambda_k\boldsymbol{D}^{(k)}$$

$$k:=k+1$$

转回第(2)步。

例 3 用可行方向法解下述非线性规划问题

$$\begin{cases}\max\ \bar{f}(\boldsymbol{X})=4x_1+4x_2-x_1^2-x_2^2\\ x_1+2x_2\leqslant 4\end{cases}$$

解 先将该非线性规划问题写成

$$\begin{cases}\min\ f(\boldsymbol{X})=-4x_1-4x_2+x_1^2+x_2^2\\ g_1(\boldsymbol{X})=-x_1-2x_2+4\geqslant 0\end{cases}$$

取初始可行点 $\boldsymbol{X}^{(0)}=(0,0)^{\mathrm{T}}$，$f(\boldsymbol{X}^{(0)})=0$

$$\nabla f(\boldsymbol{X})=\begin{pmatrix}2x_1-4\\2x_2-4\end{pmatrix},\qquad \nabla f(\boldsymbol{X}^{(0)})=\begin{pmatrix}-4\\-4\end{pmatrix}$$

$$\nabla g_1(\boldsymbol{X})=(-1,-2)^{\mathrm{T}}$$

$g_1(\boldsymbol{X}^{(0)})=4>0$，从而 $J(\boldsymbol{X}^{(0)})=\varnothing$(空集)。由于

$$\|\nabla f(\boldsymbol{X}^{(0)})\|^2=(-4)^2+(-4)^2=32$$

所以 $\boldsymbol{X}^{(0)}$ 不是(近似)极小点。现取搜索方向

$$\boldsymbol{D}^{(0)}=-\nabla f(\boldsymbol{X}^{(0)})=(4,4)^{\mathrm{T}}$$

从而

$$\boldsymbol{X}^{(1)}=\boldsymbol{X}^{(0)}+\lambda\boldsymbol{D}^{(0)}=\begin{pmatrix}0\\0\end{pmatrix}+\lambda\begin{pmatrix}4\\4\end{pmatrix}=\begin{pmatrix}4\lambda\\4\lambda\end{pmatrix}$$

将其代入约束条件，并令 $g_1(\boldsymbol{X}^{(1)})=0$，解得 $\bar{\lambda}=1/3$。

$$f(\boldsymbol{X}^{(1)})=-16\lambda-16\lambda+16\lambda^2+16\lambda^2=32\lambda^2-32\lambda$$

令 $f(\boldsymbol{X}^{(1)})$ 对 λ 的导数等于零，解得 $\lambda=1/2$。因 λ 大于 $\bar{\lambda}$ ($\bar{\lambda}=1/3$)，故取 $\lambda_0=\bar{\lambda}=1/3$。

$$\boldsymbol{X}^{(1)}=\left(\frac{4}{3},\frac{4}{3}\right)^{\mathrm{T}},\qquad f(\boldsymbol{X}^{(1)})=-\frac{64}{9}$$

$$\nabla f(\boldsymbol{X}^{(1)})=\left(-\frac{4}{3},-\frac{4}{3}\right)^{\mathrm{T}},\qquad g_1(\boldsymbol{X}^{(1)})=0$$

现构成下述线性规划问题

$$\begin{cases}\min \eta \\ -\dfrac{4}{3}d_1-\dfrac{4}{3}d_2\leqslant\eta \\ d_1+2d_2\leqslant\eta \\ -1\leqslant d_1\leqslant 1,\quad -1\leqslant d_2\leqslant 1\end{cases}$$

为便于用单纯形法求解，令

$$y_1=d_1+1,\quad y_2=d_2+1,\quad y_3=-\eta$$

从而得到

$$\begin{cases}\min(-y_3) \\ \dfrac{4}{3}y_1+\dfrac{4}{3}y_2-y_3\geqslant\dfrac{8}{3} \\ y_1+2y_2+y_3\leqslant 3 \\ y_1\leqslant 2 \\ y_2\leqslant 2 \\ y_1,y_2,y_3\geqslant 0\end{cases}$$

引入剩余变量 y_4，松弛变量 y_5、y_6 和 y_7 以及人工变量 y_8，得线性规划问题如下

$$\begin{cases}\min(-y_3+My_8) \\ \dfrac{4}{3}y_1+\dfrac{4}{3}y_2-y_3-y_4+y_8=\dfrac{8}{3} \\ y_1+2y_2+y_3+y_5=3 \\ y_1+y_6=2 \\ y_2+y_7=2 \\ y_j\geqslant 0,\quad j=1,2,\cdots,8\end{cases}$$

其最优解为

$$y_1=2,\quad y_2=3/10,\quad y_3=4/10,\quad y_4=y_5=y_6=0,\quad y_7=17/10$$

从而得到，$\eta=-y_3=-4/10$，搜索方向

$$\boldsymbol{D}^{(1)}=\begin{pmatrix}d_1\\d_2\end{pmatrix}=\begin{pmatrix}y_1-1\\y_2-1\end{pmatrix}=\begin{pmatrix}1.0\\-0.7\end{pmatrix}$$

由此

$$\boldsymbol{X}^{(2)}=\boldsymbol{X}^{(1)}+\lambda\boldsymbol{D}^{(1)}=\begin{pmatrix}4/3+\lambda\\4/3-0.7\lambda\end{pmatrix}$$

$$f(\boldsymbol{X}^{(2)})=1.49\lambda^2-0.4\lambda-7.111$$

令 $\dfrac{\mathrm{d}f(\boldsymbol{X}^{(2)})}{\mathrm{d}\lambda}=0$，得到 $\lambda=0.134$。现暂用该步长，算出

$$\boldsymbol{X}^{(2)}=\begin{pmatrix}4/3+0.134\\4/3-0.7\times 0.134\end{pmatrix}=\begin{pmatrix}1.467\\1.239\end{pmatrix}$$

因 $g_1(\boldsymbol{X}^{(2)})=0.055>0$，上面算出的 $\boldsymbol{X}^{(2)}$ 为可行点，说明选取 $\lambda_1=0.134$ 正确。

继续迭代下去，可得最优解为 $\boldsymbol{X}^*=(1.6,1.2)^{\mathrm{T}}$， $f(\boldsymbol{X}^*)=-7.2$。

原来问题的最优解不变，其目标函数值

$$\bar{f}(\boldsymbol{X}^*)=-f(\boldsymbol{X}^*)=7.2$$

8.4 制约函数法

本节介绍求解非线性规划问题的制约函数法。使用这种方法，可将非线性规划问题的求解，转化为求解一系列无约束极值问题，因而也称这种方法为序列无约束极小化技术，简记为 SUMT(sequential unconstrained minimization technique)。常用的制约函数基本上有两类：一为**惩罚函数**(或称罚函数(penalty function))；一为**障碍函数**(barrier function)，对应于这两种函数，SUMT 有**外点法**和**内点法**。

8.4.1 外点法

考虑非线性规划问题式(8-3)，为求其最优解，构造一个函数 $\psi(t)$

$$\psi(t)=\begin{cases}0, & \text{当 } t\geqslant 0\\ \infty, & \text{当 } t<0\end{cases} \tag{8-25}$$

现把 $g_j(\boldsymbol{X})$视为 t，显然

当 $\boldsymbol{X}\in R$ 时， $\psi(g_j(\boldsymbol{X}))=0,\quad j=1,2,\cdots,l;$

当 $\boldsymbol{X}\overline{\in} R$ 时， $\psi(g_i(\boldsymbol{X}))=\infty$。

再构造函数

$$\varphi(\boldsymbol{X})=f(\boldsymbol{X})+\sum_{j=1}^{l}\psi(g_j(\boldsymbol{X})) \tag{8-26}$$

现求解无约束问题

$$\min \varphi(\boldsymbol{X}) \tag{8-27}$$

若该问题有解，假定其解为 $\boldsymbol{X}^*$，则由式(8-25)应有 $\psi(g_j(\boldsymbol{X}^*))=0$。这就是说点 $\boldsymbol{X}^*\in R$。因而，$\boldsymbol{X}^*$ 不仅是问题式(8-27)的极小解，它也是原问题式(8-3)的极小解。这样一来，就把有约束问题式(8-3)的求解化成了求解无约束问题式(8-27)。

但是，用上述方法构造的函数 $\psi(t)$在 $t=0$ 处不连续，更没有导数。为此，将该函数修改为

$$\psi(t)=\begin{cases}0, & \text{当 } t\geqslant 0\\ t^2, & \text{当 } t<0\end{cases} \tag{8-28}$$

修改后的函数 $\psi(t)$，当 $t=0$ 时导数等于零，而且 $\psi(t)$和 $\psi'(t)$对任意 t 都连续。当 $X\in R$ 时仍有

$$\sum_{j=1}^{l}\psi(g_j(\boldsymbol{X}))=0$$

当 $\boldsymbol{X}\notin R$ 时

$$0<\sum_{j=1}^{l}\psi(g_j(\boldsymbol{X}))<\infty$$

我们取一个充分大的数 $M>0$，将 $\varphi(\boldsymbol{X})$ 改为

$$P(\boldsymbol{X},M)=f(\boldsymbol{X})+M\sum_{j=1}^{l}\psi(g_j(\boldsymbol{X})) \tag{8-29}$$

或等价地

$$P(\boldsymbol{X},M)=f(\boldsymbol{X})+M\sum_{j=1}^{l}[\min(0,g_j(\boldsymbol{X}))]^2 \tag{8-30}$$

从而可使 $\min P(\boldsymbol{X},M)$ 的解 $\boldsymbol{X}(M)$ 为原问题的极小解或近似极小解。若求得的 $\boldsymbol{X}(M)\in R$，则它必定是原问题的极小解。事实上，对于所有 $\boldsymbol{X}\in R$

$$\begin{aligned}f(\boldsymbol{X})+M\sum_{j=1}^{l}\psi(g_j(\boldsymbol{X}))&=P(\boldsymbol{X},M)\\&\geqslant P(\boldsymbol{X}(M),M)=f(\boldsymbol{X}(M))\end{aligned}$$

即当 $\boldsymbol{X}\in R$ 时，有 $f(\boldsymbol{X})\geqslant f(\boldsymbol{X}(M))$。

函数 $P(\boldsymbol{X},M)$ 称为**惩罚函数**，其中的第二项 $M\sum_{j=1}^{l}\psi(g_j(\boldsymbol{X}))$ 称**惩罚项**。图 8-4 示出了这种惩罚项的例子，图中左半部表示约束条件 $g(x)=x-a\geqslant 0$ 的情形，右半部则表示 $g(x)=b-x\geqslant 0$ 的情形。

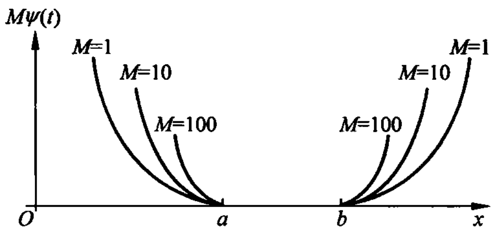


图 8-4

若对于某一个(**惩**)**罚因子** M，例如说 M_1，$\boldsymbol{X}(M_1)\notin R$，就加大罚因子的值，随着 M 值的增加，惩罚函数中的惩罚项所起的作用随之增大，$\min P(\boldsymbol{X},M)$ 的解 $\boldsymbol{X}(M)$ 与约束集 R 的“距离”就越来越近。当

$$0<M_1<M_2<\cdots<M_k<\cdots$$

趋于无穷大时，点列 $\{\boldsymbol{X}(M_k)\}$ 就从可行域 R 的外部趋于原问题式(8-3)的极小点 $\boldsymbol{X}_{\min}$（此处假设点列 $\{\boldsymbol{X}(M_k)\}$ 收敛）。

可对外点法作如下经济解释：把目标函数 $f(\boldsymbol{X})$ 看成“价格”，约束条件看成某种“规

定”,采购人可在规定范围内购置最便宜的东西。此外对违反规定制定了一种“罚款”政策,若符合规定,罚款为零;否则,要收罚款。此时,采购人付出的总代价应是价格和罚款的总和。采购者的目标是使总代价最小,这就是上述的无约束问题。当罚款规定得很苛刻时,违反规定支付的罚款很高,这就迫使采购人符合规定。在数学上表现为当罚因子 M_k 足够大时,上述无约束问题的最优解应满足约束条件,而成为约束问题的最优解。

外点法的迭代步骤如下:

(1) 取 $M_1>0$(例如说取 $M_1=1$),允许误差 $\varepsilon>0$,并令 $k:=1$。

(2) 求无约束极值问题的最优解:

$$\min_{X\in E^n} P(\boldsymbol{X},M_k)=P(\boldsymbol{X}^{(k)},M_k)$$

式中

$$P(\boldsymbol{X},M_k) = f(\boldsymbol{X}) + M_k\sum_{j=1}^{l}[\min(0,g_j(\boldsymbol{X}))]^2$$

(3) 若对某一个 $j(1\leqslant j\leqslant l)$有

$$-g_j(\boldsymbol{X}^{(k)})\geqslant\varepsilon$$

则取 $M_{k+1}>M_k$(例如,$M_{k+1}=cM_k$,$c=5$ 或 10),令 $k:=k+1$,并转向第 2 步。否则,停止迭代,得

$$\boldsymbol{X}_{\min}\approx\boldsymbol{X}^{(k)}$$

例 4 求解非线性规划

$$\begin{cases}\min f(\boldsymbol{X})=x_1+x_2\\ \quad g_1(\boldsymbol{X})=-x_1^2+x_2\geqslant 0\\ \quad g_2(\boldsymbol{X})=x_1\geqslant 0\end{cases}$$

解 构造罚函数

$$P(\boldsymbol{X},M)=x_1+x_2+M\{[\min(0,(-x_1^2+x_2))]^2+[\min(0,x_1)]^2\}$$

$$\frac{\partial P}{\partial x_1}=1+2M[\min(0,(-x_1^2+x_2)(-2x_1))]+2M[\min(0,x_1)]$$

$$\frac{\partial P}{\partial x_2}=1+2M[\min(0,(-x_1^2+x_2))]$$

对于不满足约束条件的点 $\boldsymbol{X}=(x_1,x_2)^{\mathrm{T}}$,有

$$-x_1^2+x_2<0,\quad x_1<0$$

令

$$\frac{\partial P}{\partial x_1}=\frac{\partial P}{\partial x_2}=0$$

得 $\min P(\boldsymbol{X},M)$的解为

$$\boldsymbol{X}(M)=\left(-\frac{1}{2(1+M)},\left(\frac{1}{4(1+M)^2}-\frac{1}{2M}\right)\right)^{\mathrm{T}}$$

取 $M=1,2,3,4$,可得出以下结果:

$M=1$:　　$\boldsymbol{X}=(-1/4,-7/16)^{\mathrm{T}}$

$M=2$:　　$\boldsymbol{X}=(-1/6,-2/9)^{\mathrm{T}}$

$M=3$:　　$\boldsymbol{X}=(-1/8,-29/192)^{\mathrm{T}}$

$M=4$:　　$\boldsymbol{X}=(-1/10,-23/200)^{\mathrm{T}}$

可知 $\boldsymbol{X}(M)$从 R 的外面逐步逼近 R 的边界，当 $M\to\infty$时，$\boldsymbol{X}(M)$趋于原问题的极小解 $\boldsymbol{X}_{\min}=(0,0)^{\mathrm{T}}$(见图 8-5)。

以上叙述说明，外点法的一个重要特点，就是函数 $P(\boldsymbol{X},M)$是在整个 E^n 空间内进行优化，初始点可任意选择，这给计算带来了很大方便。而且外点法也可用于非凸规划的最优化。

最后还要指出，**外点法**不只适用于含有不等式约束条件的非线性规划问题，对于等式约束条件或同时含有等式和不等式约束条件的问题也同样适用。此外，惩罚函数也可以采用其他形式。

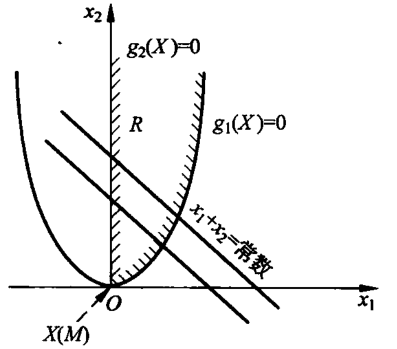


图 8-5

8.4.2 内点法

如果要求每次迭代得到的近似解都在可行域内，以便观察目标函数值的变化情况(有时可能需要这样)；或者，如果 $f(\boldsymbol{X})$在可行域外的性质比较复杂，甚至没有定义，这时就无法使用外点法。

内点法和外点法不同，它要求迭代过程始终在可行域内部进行。为此，我们把初始点取在可行域内部(即既不在可行域外，也不在可行域边界上，这种可行点称为**内点**或**严格内点**)，并在可行域的边界上设置一道“障碍”，使迭代点靠近可行域的边界时，给出的新目标函数值迅速增大，从而使迭代点始终留在可行域内部。

我们仿照外点法，通过函数叠加的办法来改造原目标函数，使得改造后的目标函数(称为**障碍函数**)具有这种性质：在可行域 R 的内部与其边界面较远的地方，障碍函数与原来的目标函数 $f(\boldsymbol{X})$尽可能相近；而在接近 R 的边界面时可以有任意大的值。可以想见，满足这种要求的障碍函数，其极小解自然不会在 R 的边界上达到。这就是说，用障碍函数来代替(近似)原目标函数，并在可行域 R 内部使其极小化，虽然 R 是一个闭集，但因极小点不在闭集的边界上，因而实际上是具有**无约束性质**的极值问题，可借助于无约束最优化的方法进行计算。

根据上述分析，即可将非线性规划式(8-3)转化为下述一系列无约束性质的极小化问题：

$$\min_{X\in R_0}\overline{P}(\boldsymbol{X},r_k) \tag{8-31}$$

其中

$$\overline{P}(\boldsymbol{X},r_k)=f(\boldsymbol{X})+r_k\sum_{j=1}^{l}\frac{1}{g_j(\boldsymbol{X})},\quad (r_k>0) \tag{8-32}$$

或

$$\overline{P}(\boldsymbol{X},r_k)=f(\boldsymbol{X})-r_k\sum_{j=1}^{l}\log(g_j(\boldsymbol{X})),\quad (r_k>0) \tag{8-33}$$

$$R_0=\{\boldsymbol{X}\mid g_j(\boldsymbol{X})>0,\quad j=1,2,\cdots,l\} \tag{8-34}$$

式(8-32)和式(8-33)右端第二项称为**障碍项**。易见，在 R 的边界上(即至少有一个 $g_j(\boldsymbol{X})=0$)，$\overline{P}(\boldsymbol{X},r_k)$为正无穷大。

如果从可行域内部的某一点 $\boldsymbol{X}^{(0)}$ 出发，按无约束极小化方法对式(8-31)进行迭代(在进行一维搜索时要**适当控制步长**，以免迭代点跑到 R_0 之外)，则随着**障碍因子** r_k 的逐步减小，即

$$r_1>r_2>\cdots>r_k>\cdots>0$$

障碍项所起的作用也越来越小，因而，求出的 $\min\overline{P}(\boldsymbol{X},r_k)$的解 $\boldsymbol{X}(r_k)$也逐步逼近原问题式(8-3)的极小解 $\boldsymbol{X}_{\min}$。若原来问题的极小解在可行域的边界上，则随着 r_k 的减小，障碍作用逐步降低，所求出的障碍函数的极小解就会不断靠近边界，直至满足某一精度要求为止。

内点法的迭代步骤如下：

(1) 取 $r_1>0$(例如取 $r_1=1$)，允许误差 $\varepsilon>0$。

(2) 找出一可行内点 $\boldsymbol{X}^{(0)}\in R_0$，并令 $k=1$。

(3) 构造障碍函数，障碍项可采用倒数函数(式(8-32))，也可采用对数函数(例如式(8-33))。

(4) 以 $\boldsymbol{X}^{(k-1)}\in R_0$ 为初始点，对障碍函数进行无约束极小化(在 R_0 内)：

$$\begin{cases}\min\limits_{x\in R_0}\overline{P}(\boldsymbol{X},r_k)=\overline{P}(\boldsymbol{X}^{(k)},r_k)\\ \qquad\boldsymbol{X}^{(k)}=\boldsymbol{X}(r_k)\in R_0\end{cases} \tag{8-35}$$

式中 $\overline{P}(\boldsymbol{X},r_k)$见式(8-32)或式(8-33)。

(5) 检验是否满足收敛准则

$$r_k\sum_{j=1}^{l}\frac{1}{g_j(\boldsymbol{X}^{(k)})}\leqslant\varepsilon$$

或

$$\left|r_k\sum_{j=1}^{l}\log(g_j(\boldsymbol{X}^{(k)}))\right|\leqslant\varepsilon$$

如满足上述准则，则以 $\boldsymbol{X}^{(k)}$ 为原问题的近似极小解 $\boldsymbol{X}_{\min}$；否则，取 $r_{k+1}<r_k$(例如取 $r_{k+1}=r_k/10$ 或 $r_k/5$)，令 $k:=k+1$，转向第(3)步继续进行迭代。

值得指出的是，根据情况，收敛准则也可采用不同的形式，例如：

$$\|\boldsymbol{X}^{(k)}-\boldsymbol{X}^{(k-1)}\|<\varepsilon$$

或

$$|f(\boldsymbol{X}^{(k)})-f(\boldsymbol{X}^{(k-1)})|<\varepsilon$$

例 5　试用内点法求解

$$\begin{cases}\min f(\boldsymbol{X})=\dfrac{1}{3}(x_1+1)^3+x_2\\ g_1(\boldsymbol{X})=x_1-1\geqslant 0\\ g_2(\boldsymbol{X})=x_2\geqslant 0\end{cases}$$

解　构造障碍函数

$$\bar{P}(\boldsymbol{X},r)=\frac{1}{3}(x_1+1)^3+x_2+\frac{r}{x_1-1}+\frac{r}{x_2}$$

$$\frac{\partial\bar{P}}{\partial x_1}=(x_1+1)^2-\frac{r}{(x_1-1)^2}=0$$

$$\frac{\partial\bar{P}}{\partial x_2}=1-\frac{r}{x_2^2}=0$$

联立解上述两个方程，得

$$x_1(r)=\sqrt{1+\sqrt{r}},\qquad x_2(r)=\sqrt{r}$$

如此得最优解：

$$\boldsymbol{X}_{\min}=\lim_{r\to 0}\left(\sqrt{1+\sqrt{r}},\sqrt{r}\right)^{\mathrm{T}}=(1,0)^{\mathrm{T}}$$

由于此例可解析求解，故可如上进行。但很多问题不便用解析法，而需用迭代法求解。

例 6　试用内点法解

$$\begin{cases}\min f(\boldsymbol{X})=x_1+x_2\\ g_1(\boldsymbol{X})=-x_1^2+x_2\geqslant 0\\ g_2(\boldsymbol{X})=x_1\geqslant 0\end{cases}$$

解　障碍项采用自然对数函数，得障碍函数如下：

$$\bar{P}(\boldsymbol{X},r)=x_1+x_2-r\log(-x_1^2+x_2)-r\log x_1$$

各次迭代结果示于表 8-1 及图 8-6 中。

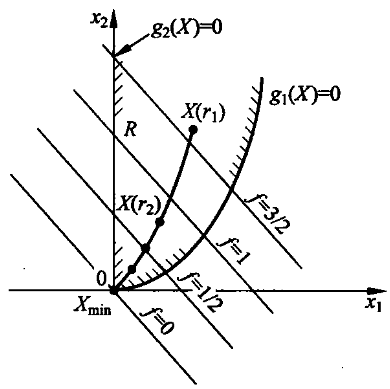


图　8-6

表　8-1

障碍因子	r	$x_1(r)$	$x_2(r)$
r_1	1.000	0.500	1.250
r_2	0.500	0.309	0.595
r_3	0.250	0.183	0.283
r_4	0.100	0.085	0.107
r_5	0.0001	0.000	0.000

我们知道,内点法的迭代过程必须由某个内点开始。在处理实际问题时,如果不能找出某个内点作为初始点,迭代就无法展开。下面说明初始内点的求法。求初始内点本身也是一个迭代过程。

先任找一点 $\boldsymbol{X}^{(0)}$ 为初始点,令

$$S_0=\{j \mid g_j(\boldsymbol{X}^{(0)})\leqslant 0,\quad 1\leqslant j\leqslant l\}$$

$$T_0=\{j \mid g_j(\boldsymbol{X}^{(0)})>0,\quad 1\leqslant j\leqslant l\}$$

如果 S_0 为空集,则 $\boldsymbol{X}^{(0)}$ 为初始内点;若 S_0 非空,则以 S_0 中的约束函数为假拟目标函数,并以 T_0 中的约束函数为障碍项,构成一无约束极值问题,对这一问题进行极小化,可得一个新点 $\boldsymbol{X}^{(1)}$。然后检验 $\boldsymbol{X}^{(1)}$ 是否为内点,若仍不为内点,如上继续进行,并减小障碍因子 r,直到求出一个内点为止。

求初始内点的迭代步骤如下:

(1) 任取一点 $\boldsymbol{X}^{(0)}\in E^n$,$r_0>0$(例如 $r_0=1$),令 $k=0$。

(2) 定出指标集 S_k 及 T_k

$$S_k=\{j \mid g_j(\boldsymbol{X}^{(k)})\leqslant 0,\quad 1\leqslant j\leqslant l\}$$

$$T_k=\{j \mid g_j(\boldsymbol{X}^{(k)})>0,\quad 1\leqslant j\leqslant l\}$$

(3) 检查集合 S_k 是否为空集,若为空集,则 $\boldsymbol{X}^{(k)}$ 在 R_0 内,初始内点找到,迭代停止。否则转向第(4)步。

(4) 构造函数

$$\widetilde{P}(\boldsymbol{X},r_k)=-\sum_{j\in S_k}g_j(\boldsymbol{X})+r_k\sum_{j\in T_k}\frac{1}{g_j(\boldsymbol{X})},\qquad r_k>0$$

以 $\boldsymbol{X}^{(k)}$ 为初始点,在保持对集合

$$\widetilde{R}_k=\{\boldsymbol{X}|_{g_j(\boldsymbol{X})>0},\qquad j\in T_k\}$$

可行的情况下,极小化 $\widetilde{P}(\boldsymbol{X},r_k)$,即

$$\min\ \widetilde{P}(\boldsymbol{X},r_k),\qquad \boldsymbol{X}\in\widetilde{R}_k$$

得 $\boldsymbol{X}^{(k+1)}$,$\boldsymbol{X}^{(k+1)}\in\widetilde{R}_k$,转向第(5)步。

(5) 令 $0<r_{k+1}<r_k$(例如说 $r_{k+1}=r_k/10$),$k:=k+1$,转向第(2)步。

注　记

关于非线性规划的系统研究始于 20 世纪 40 年代后期,1951 年 Kuhn 和 Tucker 提出了著名的 Kuhn-Tucker 条件。此后,无论在基本理论还是在实用算法的研究方面都发展很快。目前,非线性规划已成为数学规划中内容十分丰富的一个分支。限于篇幅,本篇仅叙述了非线性规划中最基本的一些概念和算法,力图便于读者掌握一些重要方法,并为进一步深入学习打好基础。

在求解无约束最优化问题的方法中,变尺度法、共轭梯度法和 Powell 直接法占有十分重要的地位。如欲进一步研究,除本篇末列出的参考文献外,还可参阅:邓乃扬等著.无约束最优化计算方法.北京:科学出版社,1982 年。

对约束极值问题，除本篇提到者外，梯度投影法、简约梯度法、约束变尺度法、乘子罚函数法、序列二次规划法和起作用约束集法(active set method)等都是很重要的方法。其中简约梯度法和起作用约束集法对处理线性约束非线性规划十分有效。处理一般非线性约束非线性规划的有效方法，有待进一步研究。

MATLAB 6.0 提供了求解不同函数的计算机程序可供使用。

希望对非线性规划有更多地了解的读者，除可参看本篇末所列文献，尤其是[8]，[12]，[13]外，还可参考：

Fletcher R. Practical Methods of Optimization, Vol. 2, John Wiley & Sons. 1981.

Edited by A. Bachem, M. Grötschel and B. Korte, Mathematical Programming: The State of the Art, Bonn 1982, Springer-Verlag, 1983.

习题

8.1 在某一试验中变更条件 x_i 四次，测得相应的结果 y_i 示于表 8-2，试为这一试验拟合一条直线，使其在最小二乘意义上最好地反映这项试验的结果(仅要求写出数学模型)。

表 8-2

x_i	2	4	6	8
y_i	1	3	5	6

8.2 有一线性方程组如下

$$\begin{cases} x_1-2x_2+3x_3=2 \\ 3x_1-2x_2+x_3=7 \\ x_1+x_2-x_3=1 \end{cases}$$

现欲用无约束极小化方法求解，试建立数学模型并说明计算原理。

8.3 试判定下述非线性规划是否为凸规划：

(1)
$$\begin{cases} \min\ f(\boldsymbol{X})=x_1^2+x_2^2+8 \\ \quad x_1^2-x_2\geqslant 0 \\ \quad -x_1-x_2^2+2=0 \\ \quad x_1, x_2\geqslant 0 \end{cases}$$

(2)
$$\begin{cases} \min\ f(\boldsymbol{X})=2x_1^2+x_2^2+x_3^2-x_1x_2 \\ \quad x_1^2+x_2^2\leqslant 4 \\ \quad 5x_1^2+x_3=10 \\ \quad x_1, x_2, x_3\geqslant 0 \end{cases}$$

8.4 试用斐波那契法求函数

$$f(x)=x^2-6x+2$$

在区间[0,10]上的极小点,要求缩短后的区间长度不大于原区间长度的8%。

8.5 试用0.618法重做习题8.4,并将计算结果与用斐波那契法所得计算结果进行比较。

8.6 试用最速下降法求解

$$\min f(\boldsymbol{X})=x_1^2+x_2^2+x_3^2$$

选初始点 $\boldsymbol{X}^{(0)}=(2,-2,1)^{\mathrm{T}}$,要求做三次迭代,并验证相邻两步的搜索方向正交。

8.7 试用最速下降法求函数

$$f(\boldsymbol{X})=-(x_1-2)^2-2x_2^2$$

的极大点。先以 $\boldsymbol{X}^{(0)}=(0,0)^{\mathrm{T}}$ 为初始点进行计算,求出极大点;再以 $\boldsymbol{X}^{(0)}=(0,1)^{\mathrm{T}}$ 为初始点进行两次迭代。最后比较从上述两个不同初始点出发的寻优过程。

8.8 试用牛顿法重解习题8.6。

8.9 试用牛顿法求解

$$\max f(\boldsymbol{X})=\frac{1}{x_1^2+x_2^2+2}$$

取初始点 $\boldsymbol{X}^{(0)}=(4,0)^{\mathrm{T}}$,用最佳步长进行。然后采用固定步长 $\lambda=1$,观察迭代情况,并加以分析说明。

8.10 试用共轭梯度法求二次函数

$$f(\boldsymbol{X})=\frac{1}{2}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}$$

的极小点,此处

$$\boldsymbol{A}=\begin{pmatrix}1 & 1\\ 1 & 2\end{pmatrix}$$

8.11 令 $\boldsymbol{X}^{(i)}(i=1,2,\cdots,n)$ 为一组 $\boldsymbol{A}$ 共轭向量(假定为列向量),$\boldsymbol{A}$ 为 $n\times n$ 对称正定阵,试证

$$\boldsymbol{A}^{-1}=\sum_{i=1}^{n}\frac{\boldsymbol{X}^{(i)}(\boldsymbol{X}^{(i)})^{\mathrm{T}}}{(\boldsymbol{X}^{(i)})^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}^{(i)}}$$

8.12 试用变尺度法求解

$$\min f(\boldsymbol{X})=(x_1-2)^3+(x_1-2x_2)^2$$

取初始点 $\boldsymbol{X}^{(0)}=(0.00,3.00)^{\mathrm{T}}$,要求近似极小点处梯度的模不大于0.5。

8.13 试以 $\boldsymbol{X}^{(0)}=(0,0)^{\mathrm{T}}$ 为初始点,使用

(1) 最速下降法(迭代4次);

(2) 牛顿法;

(3) 变尺度法。

求解无约束极值问题

$$\min\ f(\boldsymbol{X})=2x_1^2+x_2^2+2x_1x_2+x_1-x_2$$

并绘图表示使用上述各方法的寻优过程。

8.14 试用步长加速法(模矢法)求下述函数

$$f(\boldsymbol{X})=x_1^2+2x_2^2-4x_1-2x_1x_2$$

的极小点,初始点 $\boldsymbol{X}^{(0)}=(3,1)^{\mathrm{T}}$,步长

$$\boldsymbol{\Delta}_1=\begin{pmatrix}0.5\\0\end{pmatrix},\quad \boldsymbol{\Delta}_2=\begin{pmatrix}0\\0.5\end{pmatrix}$$

并绘图表示整个迭代过程。

8.15 分析非线性规划

$$\begin{cases}\min\ f(\boldsymbol{X})=(x_1-2)^2+(x_2-3)^2\\ \quad x_1^2+(x_2-2)^2\geqslant 4\\ \quad x_2\leqslant 2\end{cases}$$

在以下各点的可行下降方向(使用式(8-6)和式(8-7)):

(1) $\boldsymbol{X}^{(1)}=(0,0)^{\mathrm{T}}$;　(2) $\boldsymbol{X}^{(2)}=(2,2)^{\mathrm{T}}$;　(3) $\boldsymbol{X}^{(3)}=(3,2)^{\mathrm{T}}$。

并绘图表示各点可行下降方向的范围。

8.16 试写出下述二次规划的 Kuhn-Tucker 条件:

$$\begin{cases}\max\ f(\boldsymbol{X})=\boldsymbol{C}^{\mathrm{T}}\boldsymbol{X}+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{X}\\ \quad \boldsymbol{A}\boldsymbol{X}\leqslant \boldsymbol{b}\\ \quad \boldsymbol{X}\geqslant 0\end{cases}$$

其中:$\boldsymbol{A}$ 为 $m\times n$ 矩阵,$\boldsymbol{H}$ 为 $n\times n$ 矩阵,$\boldsymbol{C}$ 为 n 维列向量,$\boldsymbol{b}$ 为 m 维列向量,变量 $\boldsymbol{X}$ 为 n 维列向量。

8.17 试写出下述非线性规划问题的 Kuhn-Tucker 条件并进行求解

(1)
$$\begin{cases}\max\ f(x)=(x-3)^2\\ \quad 1\leqslant x\leqslant 5\end{cases}$$

(2)
$$\begin{cases}\min\ f(x)=(x-3)^2\\ \quad 1\leqslant x\leqslant 5\end{cases}$$

8.18 试找出非线性规划问题

$$\begin{cases}\max\ f(\boldsymbol{X})=x_1\\ \quad x_2-2+(x_1-1)^3\leqslant 0\\ \quad (x_1-1)^3-x_2+2\leqslant 0\\ \quad x_1,x_2\geqslant 0\end{cases}$$

的极大点,然后写出其 Kukn-Tucker 条件,这个极大点满足 Kuhn-Tucker 条件吗?试加以说明。

8.19 试解二次规划

$$\begin{cases}\min\ f(\boldsymbol{X})=2x_1^2-4x_1x_2+4x_2^2-6x_1-3x_2\\ \quad x_1+x_2\leqslant 3\\ \quad 4x_1+x_2\leqslant 9\\ \quad x_1,x_2\geqslant 0\end{cases}$$

8.20 试用可行方向法求解

$$\begin{cases}\min\ f(\boldsymbol{X})=2x_1^2+2x_2^2-2x_1x_2-4x_1-6x_2\\ \quad x_1+x_2\leqslant 2\\ \quad x_1+5x_2\leqslant 5\\ \quad x_1,x_2\geqslant 0\end{cases}$$

8.21 试用 SUMT 外点法求解

$$\begin{cases}\min\ f(\boldsymbol{X})=x_1^2+x_2^2\\ \quad x_2=1\end{cases}$$

并求出当罚因子等于 1 和 10 时的近似解。

8.22 试用 SUMT 外点法求解

$$\begin{cases}\max\ f(\boldsymbol{X})=x_1\\ \quad (x_2-2)+(x_1-1)^3\leqslant 0\\ \quad (x_1-1)^3-(x_2-2)\leqslant 0\\ \quad x_1,x_2\geqslant 0\end{cases}$$

8.23 试用 SUMT 内点法求解

$$\begin{cases}\min\ f(x)=(x+1)^2\\ \quad x\geqslant 0\end{cases}$$

8.24 试用 SUMT 内点法求解

$$\begin{cases}\min\ f(x)=x\\ \quad 0\leqslant x\leqslant 1\end{cases}$$

参考资料

[1] 南京大学数学系计算数学专业编.最优化方法.北京：科学出版社,1978
[2] 王德人编.非线性方程组解法与最优化方法.北京：人民教育出版社,1979
[3] 中国科学院数学研究所运筹室编.最优化方法.北京：科学出版社,1980
[4] 马仲蕃、魏权龄、赖炎连编.数学规划讲义.北京：中国人民大学出版社,1981
[5] 薛嘉庆编.最优化原理与方法.北京：冶金工业出版社,1983
[6] 席少霖、赵凤治编著.最优化计算方法.上海：上海科学技术出版社,1983
[7] 郭耀煌等编著.运筹学与工程系统分析.北京：中国建筑工业出版社,1986

[8] 徐光辉主编，刘彦佩、程侃副主编.运筹学基础手.北京：科学出版社，1999

[9] M.阿佛里耳著.李元熹等译.非线性规划——分析与方法.上海：上海科学技术出版社，1979

[10] D.M.希梅尔布劳著，张义燊等译.实用非线性规划.北京：科学出版社，1981

[11] D.G.鲁恩伯杰著，夏尊铨等译.线性与非线性规划引论.北京：科学出版社，1980

[12] David A Wismer，Chattergy R. Introduction To Nonlinear Optimization：A Problem Solving Approach，North-Holland Publishing Company，1978

[13] Mokhtar S Bazaraa，Shetty C M. Nonlinear Programming：Theory and Algorithms，John Wiley & Sons，1979

[14] Philip E Gill. Walter Murray and Margaret H. Wright，Practical Optimization，Academic Press，1981

Part 5 第5篇

动态规划

动态规则是运筹学的一个分支，它是解决多阶段决策过程最优化的一种数学方法。大约产生于20世纪50年代。1951年美国数学家贝尔曼(R. Bellman)等人，根据一类多阶段决策问题的特点，把多阶段决策问题变换为一系列互相联系的单阶段问题，然后逐个加以解决。与此同时，他提出了解决这类问题的"最优性原理"，研究了许多实际问题，从而创建了解决最优化问题的一种新的方法——动态规划。他的名著"动态规划"于1957年出版，该书是动态规划的第一本著作。

动态规划的方法，在工程技术、企业管理、工农业生产及军事等部门中都有广泛的应用，并且获得了显著的效果。在企业管理方面，动态规划可以用来解决最优路径问题、资源分配问题、生产调度问题、库存问题、装载问题、排序问题、设备更新问题、生产过程最优控制问题等等，所以它是现代企业管理中的一种重要的决策方法。许多问题用动态规划的方法去处理，常比线性规划或非线性规划更有成效。特别对于离散性的问题，由于解析数学无法施展其术，而动态规划的方法就成为非常有用的工具。应指出，动态规划是求解某类问题的一种方法，是考查问题的一种途径，而不是一种特殊算法(如线性规划是一种算法)。因而，它不像线性规划那样有一个标准的数学表达式和明确定义的一组规则，而必须对具体问题进行具体分析处理。因此，读者在学习时，除了要对基本概念和方法正确理解外，应以丰富的想象力去建立模型，用创造性的技巧去求解。

动态规划模型的分类，根据多阶段决策过程的时间参量是离散的还是连续的变量，过程分为离散决策过程和连续决策过程。根据决策过程的演变是确定性的还是随机性的，过程又可分为确定性决策过程和随机性决策过程。组合起来就有离散确定性、离散随机性、连续确定性、连续随机性四种决策过程模型。

本部分主要研究离散决策过程，介绍动态规划的基本概念、理论和方法，并通过几个典型的问题来说明它的应用，这些都是整个动态规划的基本内容。

CHAPTER 9 第9章

动态规划的基本方法

9.1 多阶段决策过程及实例

在生产和科学实验中，有一类活动的过程，由于它的特殊性，可将过程分为若干个互相联系的阶段，在它的每一个阶段都需要作出决策，从而使整个过程达到最好的活动效果。因此，各个阶段决策的选取不是任意确定的，它依赖于当前面临的状态，又影响以后的发展。当各个阶段决策确定后，就组成了一个决策序列，因而也就决定了整个过程的一条活动路线。这种把一个问题可看做是一个前后关联具有链状结构的多阶段过程(如图 9-1 所示)就称为多阶段决策过程，也称序贯决策过程。这种问题就称为多阶段决策过程。

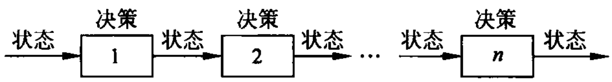


图 9-1

在多阶段决策问题中，各个阶段采取的决策，一般来说是与时间有关的，决策依赖于当前的状态，又随即引起状态的转移，一个决策序列就是在变化的状态中产生出来的，故有“动态”的含义。因此，把处理它的方法称为动态规划方法。但是，一些与时间没有关系的静态规划(如线性规划、非线性规划等)问题，只要人为地引进“时间”因素，也可把它视为多阶段决策问题，用动态规划方法去处理。

多阶段决策问题很多，现举例如下。

例 1 最短路线问题

如图 9-2，给定一个线路网络，两点之间连线上的数字表示两点间的距离(或费用)，试求一条由 A 到 G 的铺管线路，使总距离为最短(或总费用最小)。

例 2 机器负荷分配问题

某种机器可以在高低两种不同的负荷下进行生产。在高负荷下进行生产时，产品的年产量 g 和投入生产的机器数量 u_1 的关系为

$$g=g(u_1)$$

这时，机器的年完好率为 a，即如果年初完好机器的数量为 u，到年终时完好的机器就为

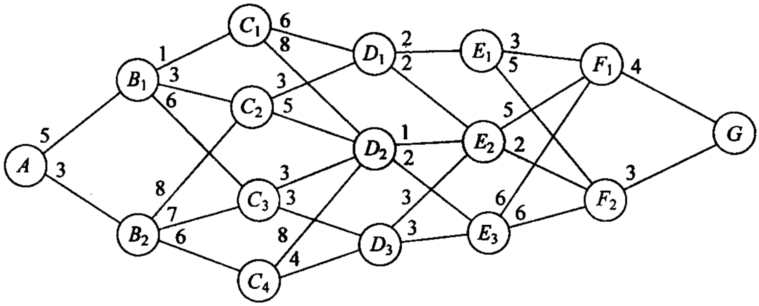


图 9-2

au,$0<a<1$,在低负荷下生产时,产品的年产量 h 和投入生产的机器数量 u_2 的关系为

$$h=h(u_2)$$

相应的机器年完好率为 b,$0<b<1$。

假定开始生产时完好的机器数量为 s_1。要求制定一个五年计划,在每年开始时,决定如何重新分配完好的机器在两种不同的负荷下生产的数量,使在五年内产品的总产量达到最高。

还有,如各种资源(人力、物力)分配问题、生产一存储问题、最优装载问题、水库优化调度问题、最优控制问题等等,都是具有多阶段决策问题的特性,均可用动态规划方法去求解。

9.2 动态规划的基本概念和基本方程

如图 9-2 所示的线路网络,求 A 到 G 的最短路线问题是动态规划中一个较为直观的典型例子。现通过讨论它的解法,来说明动态规划方法的基本思想,并阐述它的基本概念。

由图 9-2 可知,从 A 点到 G 点可以分为 6 个阶段。从 A 到 B 为第一阶段,从 B 到 C 为第二阶段…从 F 到 G 为第六阶段。在第一阶段,A 为起点,终点有 B_1、B_2 两个,因而这时走的路线有两个选择,一是走到 B_1;一是走到 B_2,若选择走到 B_2 的决策,则 B_2 就是第一阶段在我们决策之下的结果。它既是第一阶段路线的终点,又是第二阶段路线的始点。在第二阶段,再从 B_2 点出发,对应于 B_2 点就有一个可供选择的终点集合$\{C_2,C_3,C_4\}$;若选择由 B_2 走至 C_2 为第二阶段的决策,则 C_2 就是第二阶段的终点,同时又是第三阶段的始点。同理递推下去,可看到:各个阶段的决策不同,铺管路线就不同。很明显,当某阶段的始点给定时,它直接影响着后面各阶段的行进路线和整个路线的长短,而后面各阶段的路线的发展不受这点以前各阶段路线的影响。故此问题的要求是:在各个阶段选取一个恰当的决策,使由这些决策组成的一个决策序列所决定的一条路线,其总路程最短。

如何解决这个问题呢？可以采取穷举法。即把由 A 到 G 所有可能的每一条路线的距离都算出来，然后互相比较找出最短者，相应地得出了最短路线。这样，由 A 到 G 的 6 个阶段中，一共有 $2\times3\times2\times2\times2\times1=48$ 条不同的路线，比较 48 条不同的路线的距离值，才能找出最短路线为

$$A\to B_1\to C_2\to D_1\to E_2\to F_2\to G$$

相应最短距离为 18。显然，这样作计算是相当繁杂的。如果当段数很多，各段的不同选择也很多时，这种解法的计算将变得极其繁杂，甚至在电子计算机上计算都是不现实的。因此，为了减少计算工作量，需要寻求更好的算法，这就是下面要介绍的动态规划的方法。为了讨论方便，先介绍动态规划的基本概念和符号。

9.2.1 动态规划的基本概念

1. 阶段

把所给问题的过程，恰当地分为若干个相互联系的阶段，以便能按一定的次序去求解。描述阶段的变量称为阶段变量，常用 k 表示。阶段的划分，一般是根据时间和空间的自然特征来划分，但要便于把问题的过程能转化为多阶段决策的过程。如例 1 可分为 6 个阶段来求解，k 分别等于 1、2、3、4、5、6。

2. 状态

状态表示每个阶段开始所处的自然状况或客观条件，它描述了研究问题过程的状况，又称不可控因素。在例 1 中，状态就是某阶段的出发位置。它既是该阶段某支路的起点，又是前一阶段某支路的终点。通常一个阶段有若干个状态，第一阶段有一个状态就是点 A，第二阶段有两个状态，即点集合$\{B_1,B_2\}$，一般第 k 阶段的状态就是第 k 阶段所有始点的集合。

描述过程状态的变量称为状态变量。它可用一个数、一组数或一向量（多维情形）来描述。常用 S_k 表示第 k 阶段的状态变量。如在例 1 中第三阶段有四个状态，则状态变量 S_k 可取四个值，即 C_1、C_2、C_3、C_4。点集合$\{C_1,C_2,C_3,C_4\}$就称为第三阶段的**可达状态集合**。记为 $S_3=\{C_1,C_2,C_3,C_4\}$。有时为了方便起见，将该阶段的状态编上号码 1，2…这时也可记 $S_3=\{1,2,3,4\}$。第 k 阶段的可达状态集合就记为 S_k。

这里所说的状态应具有下面的性质：如果某阶段状态给定后，则在这阶段以后过程的发展不受这阶段以前各段状态的影响。换句话说，过程的过去历史只能通过当前的状态去影响它未来的发展，当前的状态是以往历史的一个总结。这个性质称为**无后效性**（即马尔科夫性）。

如果状态仅仅描述过程的具体特征，则并不是任何实际过程都能满足无后效性的要求。所以，在构造决策过程的动态规划模型时，不能仅由描述过程的具体特征这点着眼去规定状态变量，而要充分注意是否满足无后效性的要求。如果状态的某种规定方式可能

导致不满足无后效性，应适当地改变状态的规定方法，达到能使它满足无后效性的要求。例如，研究物体(把它看做一个质点)受外力作用后其空间运动的轨迹问题。从描述轨迹这点着眼，可以只选坐标位置(x_k, y_k, z_k)作为过程的状态，但这样不能满足无后效性，因为即使知道了外力的大小和方向，仍无法确定物体受力后的运动方向和轨迹，只有把位置(x_k, y_k, z_k)和速度$(\dot{x}_k, \dot{y}_k, \dot{z}_k)$都作为过程的状态变量，才能确定物体运动下一步的方向和轨迹，实现无后效性的要求。

3. 决策

决策表示当过程处于某一阶段的某个状态时，可以作出不同的决定(或选择)，从而确定下一阶段的状态，这种决定称为决策。在最优控制中也称为控制。描述决策的变量，称为决策变量。它可用一个数、一组数或一向量来描述。常用$u_k(s_k)$表示第k阶段当状态处于s_k时的决策变量。它是状态变量的函数。在实际问题中，决策变量的取值往往限制在某一范围之内，此范围称为**允许决策集合**。常用$D_k(s_k)$表示第k阶段从状态s_k出发的允许决策集合，显然有$u_k(s_k) \in D_k(s_k)$。

如在例1第二阶段中，若从状态B_1出发，就可作出三种不同的决策，其允许决策集合$D_2(B_1)=\{C_1, C_2, C_3\}$，若选取的点为$C_2$，则$C_2$是状态$B_1$在决策$u_2(B_1)$作用下的一个新的状态，记作$u_2(B_1)=C_2$。

4. 策略

策略是一个按顺序排列的决策组成的集合。由过程的第k阶段开始到终止状态为止的过程，称为问题的后部子过程(或称为k子过程)。由每段的决策按顺序排列组成的决策函数序列$\{u_k(s_k), \cdots, u_n(s_n)\}$称为$k$子过程策略，简称子策略，记为$p_{k,n}(s_k)$。即

$$p_{k,n}(s_k)=\{u_k(s_k), u_{k+1}(s_{k+1}), \cdots, u_n(s_n)\}$$

当$k=1$时，此决策函数序列称为全过程的一个策略，简称策略，记为$p_{1,n}(s_1)$。即

$$p_{1,n}(s_1)=\{u_1(s_1), u_2(s_2), \cdots, u_n(s_n)\}$$

在实际问题中，可供选择的策略有一定的范围，此范围称为**允许策略集合**，用P表示。从允许策略集合中找出达到最优效果的策略称为**最优策略**。

5. 状态转移方程

状态转移方程是确定过程由一个状态到另一个状态的演变过程。若给定第k阶段状态变量s_k的值，如果该段的决策变量u_k一经确定，第$k+1$阶段的状态变量s_{k+1}的值也就完全确定。即s_{k+1}的值随s_k和u_k的值变化而变化。这种确定的对应关系，记为

$$s_{k+1}=T_k(s_k, u_k)$$

上式描述了由k阶段到$k+1$阶段的状态转移规律，称为状态转移方程。T_k称为状态转移函数。如例1中，状态转移方程为$s_{k+1}=u_k(s_k)$。

6. 指标函数和最优值函数

用来衡量所实现过程优劣的一种数量指标，称为指标函数。它是定义在全过程和所有后部子过程上确定的数量函数。常用$V_{k,n}$表示之。即

$$V_{k,n}=V_{k,n}(s_k,u_k,s_{k+1},\cdots,s_{n+1}),k=1,2,\cdots,n$$

对于要构成动态规划模型的指标函数，应具有可分离性，并满足递推关系。即 $V_{k,n}$ 可以表示为 s_k、u_k、$V_{k+1,n}$ 的函数。记为

$$V_{k,n}(s_k,u_k,s_{k+1},\cdots,s_{n+1})=\psi_k[s_k,u_k,V_{k+1,n}(s_{k+1},\cdots,s_{n+1})]$$

在实际问题中很多指标函数都满足这个性质。

常见的指标函数的形式如下。

(1) 过程和它的任一子过程的指标是它所包含的各阶段的指标的和。即

$$V_{k,n}(s_k,u_k,\cdots,s_{n+1})=\sum_{j=k}^{n}v_j(s_j,u_j)$$

式中 $v_j(s_j,u_j)$ 表示第 j 阶段的阶段指标。这时上式可写成

$$V_{k,n}(s_k,u_k,\cdots,s_{n+1})=v_k(s_k,u_k)+V_{k+1,n}(s_{k+1},u_{k+1},\cdots,s_{n+1})$$

(2) 过程和它的任一子过程的指标是它所包含的各阶段的指标的乘积。即

$$V_{k,n}(s_k,u_k,\cdots,s_{n+1})=\prod_{j=k}^{n}v_j(s_j,u_j)$$

这时就可写成

$$V_{k,n}(s_k,u_k,\cdots,s_{n+1})=v_k(s_k,u_k)V_{k+1,n}(s_{k+1},u_{k+1},\cdots,s_{n+1})$$

指标函数的最优值，称为**最优值函数**，记为 $f_k(s_k)$。它表示从第 k 阶段的状态 s_k 开始到第 n 阶段的终止状态的过程，采取最优策略所得到的指标函数值。即

$$f_k(s_k)=\underset{\{u_k,\cdots,u_n\}}{\text{opt}}V_{k,n}(s_k,u_k,\cdots,s_{n+1})$$

其中"opt"是最优化(optimization)的缩写，可根据题意而取 min 或 max。

在不同的问题中，指标函数的含义是不同的，它可能是距离、利润、成本、产品的产量或资源消耗等。例如，在最短路线问题中，指标函数 $V_{k,n}$ 就表示在第 k 阶段由点 s_k 至终点 G 的距离。用 $d_k(s_k,u_k)=v_k(s_k,u_k)$ 表示在第 k 阶段由点 s_k 到点 $s_{k+1}=u_k(s_k)$ 的距离，如 $d_5(E_1,F_1)=3$，就表示在第 5 阶段中由点 E_1 到点 F_1 的距离为 3。$f_k(s_k)$ 表示从第 k 阶段点 s_k 到终点 G 的最短距离，如 $f_4(D_1)$ 就表示从第 4 阶段中的点 D_1 到点 G 的最短距离。

9.2.2 动态规划的基本思想和基本方程

现在，再结合解决最短路线问题来介绍动态规划方法的基本思想。生活中的常识告诉我们，最短路线有一个重要特性：如果由起点 A 经过 P 点和 H 点而到达终点 G 是一条最短路线，则由点 P 出发经过 H 点到达终点 G 的这条子路线，对于从点 P 出发到达终点的所有可能选择的不同路线来说，必定也是最短路线。例如，在最短路线问题中，若找到了 $A\to B_1\to C_2\to D_1\to E_2\to F_2\to G$ 是由 A 到 G 的最短路线，则 $D_1\to E_2\to F_2\to G$ 应该是由 D_1 出发到 G 点的所有可能选择的不同路线中的最短路线。此特性用反证法易证。因为如果不是这样，则从点 P 到 G 点有另一条距离更短的路线存在，把它和原来最短路线

由 A 点到达 P 点的那部分连接起来，就会得到一条由 A 点到 G 点的新路线，它比原来那条最短路线的距离还要短些。这与假设矛盾，是不可能的。

根据最短路线这一特性，寻找最短路线的方法，就是从最后一段开始，用由后向前逐步递推的方法，求出各点到 G 点的最短路线，最后求得由 A 点到 G 点的最短路线。所以，动态规划的方法是从终点逐段向始点方向寻找最短路线的一种方法，如图 9-3 所示。

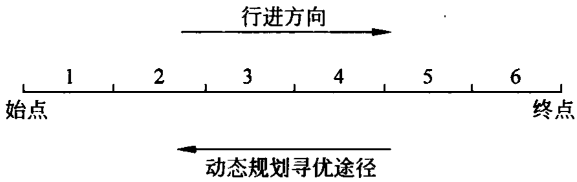


图 9-3

下面按照动态规划的方法，将例 1 从最后一段开始计算，由后向前逐步推移至 A 点。

当 $k=6$ 时，由 F_1 到终点 G 只有一条路线，故 $f_6(F_1)=4$。同理，$f_6(F_2)=3$。

当 $k=5$ 时，出发点有 E_1、E_2、E_3 三个。若从 E_1 出发，则有两个选择：①至 F_1；②至 F_2

则

$$f_5(E_1)=\min\begin{Bmatrix}d_5(E_1,F_1)+f_6(F_1)\\d_5(E_1,F_2)+f_6(F_2)\end{Bmatrix}=\min\begin{Bmatrix}3+4\\5+3\end{Bmatrix}=7$$

其相应的决策为 $u_5(E_1)=F_1$

这说明，由 E_1 至终点 G 的最短距离为 7，其最短路线是

$$E_1\to F_1\to G$$

同理，从 E_2 和 E_3 出发，则有

$$f_5(E_2)=\min\begin{Bmatrix}d_5(E_2,F_1)+f_6(F_1)\\d_5(E_2,F_2)+f_6(F_2)\end{Bmatrix}=\min\begin{Bmatrix}5+4\\2+3\end{Bmatrix}=5$$

其相应的决策为 $u_5(E_2)=F_2$

$$f_5(E_3)=\min\begin{Bmatrix}d_5(E_3,F_1)+f_6(F_1)\\d_5(E_3,F_2)+f_6(F_2)\end{Bmatrix}=\min\begin{Bmatrix}6+4\\6+3\end{Bmatrix}=9$$

且 $u_5(E_3)=F_2$

类似地，可算得

当 $k=4$ 时，有

$$f_4(D_1)=7\qquad u_4(D_1)=E_2$$
$$f_4(D_2)=6\qquad u_4(D_2)=E_2$$
$$f_4(D_3)=8\qquad u_4(D_3)=E_2$$

当 $k=3$ 时，有

$$f_3(C_1)=13 \qquad u_3(C_1)=D_1$$
$$f_3(C_2)=10 \qquad u_3(C_2)=D_1$$
$$f_3(C_3)=9 \qquad u_3(C_3)=D_2$$
$$f_3(C_4)=12 \qquad u_3(C_4)=D_3$$

当 $k=2$ 时，有

$$f_2(B_1)=13 \qquad u_2(B_1)=C_2$$
$$f_2(B_2)=16 \qquad u_2(B_2)=C_3$$

当 $k=1$ 时，出发点只有一个 A 点，则

$$f_1(A)=\min\begin{Bmatrix}d_1(A,B_1)+f_2(B_1)\\d_1(A,B_2)+f_2(B_2)\end{Bmatrix}=\min\begin{Bmatrix}5+13\\3+16\end{Bmatrix}=18$$

且 $u_1(A)=B_1$。于是得到从起点 A 到终点 G 的最短距离为 18。

为了找出最短路线，再按计算的顺序反推之，可求出最优决策函数序列 $\{u_k\}$，即由 $u_1(A)=B_1, u_2(B_1)=C_2, u_3(C_2)=D_1, u_4(D_1)=E_2, u_5(E_2)=F_2, u_6(F_2)=G$ 组成一个最优策略。因而，找出相应的最短路线为

$$A\to B_1\to C_2\to D_1\to E_2\to F_2\to G$$

从上面的计算过程中可以看出，在求解的各个阶段，我们利用了 k 阶段与 $k+1$ 阶段之间的递推关系：

$$\begin{cases}f_k(s_k)=\min\limits_{u_k\in D_k(s_k)}\{d_k(s_k,u_k(s_k))+f_{k+1}(u_k(s_k))\} \quad k=6,5,4,3,2,1\\ f_7(s_7)=0(\text{或写成 } f_6(s_6)=d_6(s_6,G))\end{cases}$$

一般情况，k 阶段与 $k+1$ 阶段的递推关系式可写为

$$f_k(s_k)=\underset{u_k\in D_k(s_k)}{\operatorname{opt}}\{v_k(s_k,u_k(s_k))+f_{k+1}(u_k(s_k))\} \tag{9-1}$$
$$k=n,n-1,\cdots,1$$

边界条件为

$$f_{n+1}(s_{n+1})=0$$

这种递推关系式(9-1)称为**动态规划的基本方程**。

现在把动态规划方法的基本思想归纳如下。

(1) 动态规划方法的关键在于正确地写出基本的递推关系式和恰当的边界条件(简言之为基本方程)。要做到这一点，必须先将问题的过程分成几个相互联系的阶段，恰当地选取状态变量和决策变量及定义最优值函数，从而把一个大问题化成一族同类型的子问题，然后逐个求解。即从边界条件开始，逐段递推寻优，在每一个子问题的求解中，均利用了它前面的子问题的最优化结果，依次进行，最后一个子问题所得的最优解，就是整个问题的最优解。

(2) 在多阶段决策过程中，动态规划方法是既把当前一段和未来各段分开，又把当前效益和未来效益结合起来考虑的一种最优化方法。因此，每段决策的选取是从全局来考

虑的，与该段的最优选择答案一般是不同的。

(3) 在求整个问题的最优策略时，由于初始状态是已知的，而每段的决策都是该段状态的函数，故最优策略所经过的各段状态便可逐次变换得到，从而确定了最优路线。

如例 1 最短路线问题，初始状态 A 已知，则按下面箭头所指的方向逐次变换有

$$\begin{array}{ccccccccccc} & & u_1(A) & & u_2(B_1) & & \cdots & & u_6(F_2) & & \\ & \nearrow & & \downarrow\ \nearrow & & \downarrow\ \nearrow & & \nearrow & & \downarrow & \\ A & & & B_1 & & C_2 & \cdots & & & G & \end{array}$$

(已知)

从而可得最优策略为 $\{u_1(A), u_2(B_1), \cdots, u_6(F_2)\}$，相应的最短路线为

$$A \to B_1 \to C_2 \to D_1 \to E_2 \to F_2 \to G$$

上述最短路线问题的计算过程，也可借助图形直观简明地表示出来，如图 9-4 所示。

在图 9-4 中，每节点处上方的方格内的数，表示该点到终点 G 的最短距离。用直线连接的点表示该点到终点 G 的最短路线。未用直线连接的点就说明它不是该点到终点 G 的最短路线，故这些支路均被舍去了。图中粗线表示由始点 A 到终点 G 的最短路线。

这种在图上直接作业的方法叫做**标号法**。如果规定从 A 点到 G 点为顺行方向，则由 G 点到 A 点为逆行方向，那么，图 9-4 是由 G 点开始从后向前标的。这种以 A 为始端，G 为终端，从 G 到 A 的解法称为**逆序解法**。

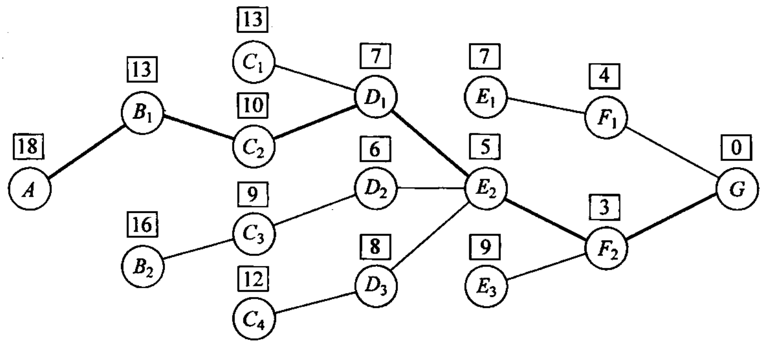


图 9-4

由于线路网络的两端都是固定的，且线路上的数字是表示两点间的距离，则从 A 点计算到 G 点和从 G 点计算到 A 点的最短路线是相同的。因而，标号也可以由 A 开始，从前向后标。只是那时是视 G 为起点，A 为终点，按动态规划方法处理的，如图 9-5所示。

图 9-5 中，每节点处上方方格内的数表示该点到 A 点的最短距离，用直线连接的点表示该点到起点 A 的最短路线，粗线表示 A 到 G 的最短路线。

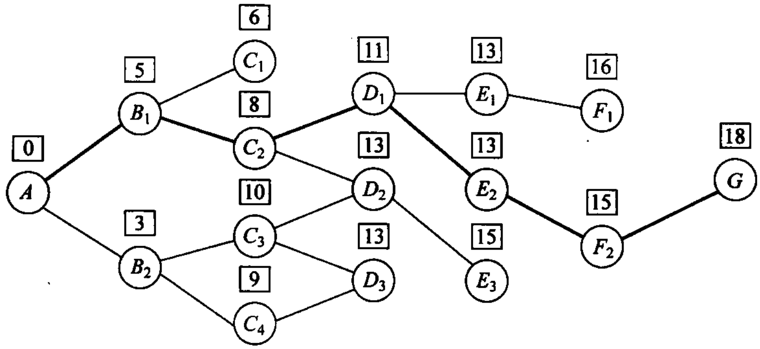


图 9-5

这种以 A 为始端、G 为终端的从 A 到 G 的解法称为**顺序解法**。

由此可见,顺序解法和逆序解法只表示行进方向的不同或对始端终端看法的颠倒。但用动态规划方法求最优解时,都是在行进方向规定后,均要逆着这个规定的行进方向,从最后一段向前逆推计算,逐段找出最优途径。

从上面例 1 的计算过程,明显看到,动态规划的方法比穷举法有以下优点:

(1) 减少了计算量。计算例 1 若用穷举法,就要对 48 条路线进行比较,运算在计算机上进行时,比较运算要进行 47 次;求各条路线的距离,即使用逐段累加方法,也要进行 6+12+24+48+48=138 次加法运算。

用动态规划方法来计算,比较运算(从 $k=5$ 段开始向前算)共进行 3+3+4+4+1=15 次。每次比较运算相应有两次加法运算,再去掉中间重复两次(即 $B_1 \to C_1$,$B_2 \to C_4$ 各多算了一次),实际只有 28 次加法运算。可见,动态规划方法比穷举法减少了计算量。而且随着段数的增加,计算量将大大地减少。

(2) 丰富了计算结果。在逆序(或顺序)解法中,我们得到的不仅仅是由 A 点(或 G 点)出发到 G 点(或 A 点)的最短路线及相应的最短距离,而且得到了从所有各中间点出发到 G 点(或 A 点)的最短路线及相应的距离。这就是说,求出的不是一个最优策略,而是一族的最优策略。这对许多实际问题来讲是很有用的,有利于帮助分析所得结果。

在明确了动态规划的基本概念和基本思想之后,我们看到,给一个实际问题建立动态规划模型时,必须做到下面五点:

(1) 将问题的过程划分成恰当的阶段;

(2) 正确选择状态变量 s_k,使它既能描述过程的演变,又要满足无后效性;

(3) 确定决策变量 u_k 及每阶段的允许决策集合 $D_k(s_k)$;

(4) 正确写出状态转移方程;

(5) 正确写出指标函数 $V_{k,n}$ 的关系,它应满足下面三个性质:

① 是定义在全过程和所有后部子过程上的数量函数;

② 要具有可分离性，并满足递推关系。即

$$V_{k,n}(s_k,u_k,\cdots,s_{n+1})=\psi_k[s_k,u_k,V_{k+1,n}(s_{k+1},u_{k+1},\cdots,s_{n+1})]$$

③ 函数 $\psi_k(s_k,u_k,V_{k+1,n})$ 对于变量 $V_{k+1,n}$ 要严格单调。

以上五点是构造动态规划模型的基础，是正确写出动态规划基本方程的基本要素。而一个问题的动态规划模型是否正确给出，它集中地反映在恰当的定义最优值函数和正确地写出递推关系式及边界条件上。简言之，要正确写出动态规划的基本方程。

根据动态规划方法有逆序解法和顺序解法之分，那么，它们的动态规划基本方程应如何来表述呢？

设指标函数是取各阶段指标的和的形式，即

$$V_{k,n}=\sum_{j=k}^{n}v_j(s_j,u_j)$$

其中 $v_j(s_j,u_j)$ 表示第 j 段的指标。它显然是满足指标函数三个性质的。所以上式可写成

$$V_{k,n}=v_k(s_k,u_k)+V_{k+1,n}[s_{k+1},\cdots,s_{n+1}]$$

当初始状态给定时，过程的策略就被确定，则指标函数也就确定了。因此，指标函数是初始状态和策略的函数。可记为 $V_{k,n}[s_k,p_{k,n}(s_k)]$。故上面递推关系又可写成

$$V_{k,n}[s_k,p_{k,n}]=v_k(s_k,u_k)+V_{k+1,n}[s_{k+1},p_{k+1,n}]$$

其子策略 $p_{k,n}(s_k)$ 可看成是由决策 $u_k(s_k)$ 和 $p_{k+1,n}(s_{k+1})$ 组合而成。即

$$p_{k,n}=\{u_k(s_k),p_{k+1,n}(s_{k+1})\}$$

如果用 $p_{k,n}^*(s_k)$ 表示初始状态为 s_k 的后部子过程所有子策略中的最优子策略，则最优值函数为

$$f_k(s_k)=V_{k,n}[s_k,p_{k,n}^*(s_k)]=\underset{p_{k,n}}{\text{opt}}\{V_{k,n}[s_k,p_{k,n}(s_k)]\}$$

而

$$\begin{aligned}\underset{p_{k,n}}{\text{opt}}\{V_{k,n}(s_k,p_{k,n})\}&=\underset{\{u_k,p_{k+1,n}\}}{\text{opt}}\{v_k(s_k,u_k)+V_{k+1,n}(s_{k+1},p_{k+1,n})\}\\&=\underset{u_k}{\text{opt}}\{v_k(s_k,u_k)+\underset{p_{k+1,n}}{\text{opt}}V_{k+1,n}\}\end{aligned}$$

但

$$f_{k+1}(s_{k+1})=\underset{p_{k+1,n}}{\text{opt}}\{V_{k+1,n}(s_{k+1},p_{k+1,n})\}$$

所以

$$f_k(s_k)=\underset{u_k\in D_k(s_k)}{\text{opt}}\{[v_k(s_k,u_k)+f_{k+1}(s_{k+1})]\}\quad k=n,n-1,\cdots,1$$

边界条件为 $f_{n+1}(s_{n+1})=0$。

这就是**动态规划逆序解法的基本方程**。式中 $s_{k+1}=T_k(s_k,u_k)$，其求解过程，根据边界条件，从 $k=n$ 开始，由后向前逆推，从而逐步可求得各段的最优决策和相应的最优值，最后求出 $f_1(s_1)$ 时，就得到整个问题的最优解。

对于动态规划顺序解法的基本方程应如何表述呢？假定阶段序数 k 和状态变量 s_k 的定义不变，而改变决策变量 u_k 的定义，如例 1 中取 $u_k(s_{k+1})=s_k$，则这时的状态转移不是由 s_k、u_k 去确定 s_{k+1}，而是反过来由 s_{k+1}、u_k 去确定 s_k，则状态转移方程一般形式为

$$s_k=T_k^r(s_{k+1},u_k)$$

因而第 k 阶段的允许决策集合也应作相应的改变，记为 $D_k^r(s_{k+1})$。指标函数也应换成以 s_{k+1} 和 u_k 的函数表示。于是可得**动态规划顺序解法的基本方程**为

$$f_k(s_{k+1})=\underset{u_k\in D_k^r(s_{k+1})}{\text{opt}}\{v_k(s_{k+1},u_k)+f_{k-1}(s_k)\}\quad k=1,2,\cdots,n$$

边界条件为 $\qquad f_0(s_1)=0$

式中 $s_k=T_k^r(s_{k+1},u_k)$。其求解过程：根据边界条件，从 $k=1$ 开始，由前向后顺推，逐步求得各段的最优决策和相应的最优值，最后求出 $f_n(s_{n+1})$，就得到整个问题的最优解。

9.3 动态规划的最优性原理和最优性定理

20 世纪 50 年代，R. Bellman 等人根据研究一类多阶段决策问题，提出了最优性原理（有的翻译成最优化原理）作为动态规划的理论基础。用它去解决许多决策过程的优化问题。长期以来，许多动态规划的著作都用“依据最优化原理，则有……”的提法去处理决策过程的优化问题。人们对用这样的一个简单的原理作为动态规划方法的理论根据很难理解。的确如此，实际上，这提法是给用动态规划方法去处理决策过程的优化问题披上神秘的色彩，使读者不能正确理解动态规划方法的本质。

下面将介绍“动态规划的最优性原理”的原文含义。并指出它为什么不是动态规划的理论基础，进而揭示动态规划方法的本质。其理论基础是“最优性定理”。

动态规划的最优性原理：“作为整个过程的最优策略具有这样的性质：即无论过去的状态和决策如何，对前面的决策所形成的状态而言，余下的诸决策必须构成最优策略。”简言之，一个最优策略的子策略总是最优的。

但是，随着人们深入地研究动态规划，逐渐认识到：对于不同类型的问题所建立严格定义的动态规划模型，必须对相应的最优性原理给以必要的验证。就是说，最优性原理不是对任何决策过程都普遍成立的；而且“最优性原理”与动态规划基本方程，并不是无条件等价的，两者之间也不存在确定的蕴含关系。可见动态规划的基本方程在动态规划的理论和方法中起着非常重要作用。而反映动态规划基本方程的是最优性定理，它是策略最优性的充分必要条件，而最优性原理仅仅是策略最优性的必要条件，它是最优性定理的推论。在求解最优策略时，更需要的是其充分条件。所以，动态规划的基本方程或者说最优性定理才是动态规划的理论基础。

动态规划的最优性定理：设阶段数为 n 的多阶段决策过程，其阶段编号为

$$k=0,1,\cdots,n-1$$

允许策略 $p^*_{0,n-1}=(u^*_0,u^*_1,\cdots,u^*_{n-1})$为最优策略的充要条件是对任意一个 k

$0<k<n-1$ 和 $s_0\in S_0$ 有

$$V_{0,n-1}(s_0,p^*_{0,n-1})=\underset{p_{0,k-1}\in P_{0,k-1}(s_0)}{\text{opt}}\{V_{0,k-1}(s_0,p_{0,k-1})+\underset{p_{k,n-1}\in P_{k,n-1}(\tilde{s}_k)}{\text{opt}}V_{k,n-1}(\tilde{s}_k,p_{k,n-1})\}\tag{9-2}$$

式中 $p^*_{0,n-1}=(p_{0,k-1},p_{k,n-1})$，$\tilde{s}_k=T_{k-1}(s_{k-1},u_{k-1})$，它是由给定的初始状态 s_0 和子策略 $p_{0,k-1}$所确定的 k 段状态。当 V 是效益函数时，opt 取 max；当 V 是损失函数时，opt 取 min。

证明 必要性：设 $p^*_{0,n-1}$是最优策略，则

$$V_{0,n-1}(s_0,p^*_{0,n-1})=\underset{p_{0,n-1}\in P_{0,n-1}(s_0)}{\text{opt}}\{V_{0,n-1}(s_0,p_{0,n-1})\}$$
$$=\underset{p_{0,n-1}\in P_{0,n-1}(s_0)}{\text{opt}}\{[V_{0,k-1}(s_0,p_{0,k-1})+V_{k,n-1}(\tilde{s}_k,p_{k,n-1})]\}$$

但对于从 k 至 $n-1$ 阶段的子过程而言，它的总指标取决于过程的起始点

$\tilde{s}_k=T_{k-1}(s_{k-1},u_{k-1})$和子策略 $p_{k,n-1}$

而这个起始点 $\tilde{s}_k$ 是由前一段子过程在子策略 $p_{0,k-1}$下而确定的。

因此，在策略集合 $p_{0,n-1}$上求最优解，就等价于先在子策略集合 $p_{k,n-1}(\tilde{s}_k)$上求最优解，然后再求这些子最优解在子策略集合 $p_{0,k-1}(s_0)$上的最优解。故上式可写为

$$V_{0,n-1}(s_0,p^*_{0,n-1})=\underset{p_{0,k-1}\in P_{0,k-1}(s_0)}{\text{opt}}\{\underset{p_{k,n-1}\in P_{k,n-1}(\tilde{s}_k)}{\text{opt}}[V_{0,k-1}(s_0,p_{0,k-1})+V_{k,n-1}(\tilde{s}_k,p_{k,n-1})]\}$$

但括号内第一项与子策略 $p_{k,n-1}$无关，故得

$$V_{0,n-1}(s_0,p^*_{0,n-1})=\underset{p_{0,k-1}\in P_{0,k-1}(s_0)}{\text{opt}}\{V_{0,k-1}(s_0,p_{0,k-1})+\underset{p_{k,n-1}\in P_{k,n-1}(s_0)}{\text{opt}}V_{k,n-1}(\tilde{s}_k,p_{k,n-1})\}$$

充分性：设 $p_{0,n-1}=(p_{0,k-1},p_{k,n-1})$为任一策略，$\tilde{s}_k$ 为由$(s_0,p_{0,k-1})$所确定的 k 阶段的起始状态。则有

$$V_{k,n-1}(\tilde{s}_k,p_{k,n-1})\preccurlyeq\underset{p_{k,n-1}\in P_{k,n-1}(\tilde{s}_k)}{\text{opt}}\{V_{k,n-1}(\tilde{s}_k,p_{k,n-1})\}$$

在这里记号“$\preccurlyeq$”的含义是：当 opt 表示 max 时就表示“$\leqslant$”，当 opt 表示 min 时就表示“$\geqslant$”。因此，式(9-2)为

$$\begin{aligned}V_{0,n-1}(s_0,p_{0,n-1})&=V_{0,k-1}(s_0,p_{0,k-1})+V_{k,n-1}(\tilde{s}_k,p_{k,n-1})\\&\preccurlyeq V_{0,k-1}(s_0,p_{0,k-1})+\underset{p_{k,n-1}\in P_{k,n-1}(\tilde{s}_k)}{\text{opt}}\{V_{k,n-1}(\tilde{s}_k,p_{k,n-1})\}\\&\preccurlyeq\underset{p_{0,k-1}\in P_{0,k-1}(s_0)}{\text{opt}}\{V_{0,k-1}(s_0,p_{0,k-1})+\\&\quad\underset{p_{k,n-1}\in P_{k,n-1}(\tilde{s}_k)}{\text{opt}}V_{k,n-1}(\tilde{s}_k,p_{k,n-1})\}\\&=V_{0,n-1}(s_0,p^*_{0,n-1})\end{aligned}$$

故只要 $p^*_{0,n-1}$使式(9-2)成立，则对任一策略 $p_{0,n-1}$，都有

$$V_{0,n-1}(s_0, p_{0,n-1}) \leqslant V_{0,n-1}(s_0, p_{0,n-1}^*)$$

因为，$p_{0,n-1}^*$是最优策略。证毕。

推论 若允许策略 $p_{0,n-1}^*$是最优策略，则对任意的 k，$0<k<n-1$，它的子策略 $p_{k,n-1}^*$对于以 $s_k^*=T_{k-1}(s_{k-1}^*, u_{k-1}^*)$为起点的 k 到 $n-1$ 子过程来说，必是最优策略（注意：k 阶段状态 s_k^* 是由 s_0 和 $p_{0,k-1}^*$所确定的）。

证明 用反证法。

若 $p_{k,n-1}^*$不是最优策略，则有

$$V_{k,n-1}(s_k^*, p_{k,n-1}^*) \prec \operatorname*{opt}_{p_{k,n-1} \in P_{k,n-1}(s_k^*)} \{V_{k,n-1}(s_k^*, p_{k,n-1})\}$$

此处记号“$\prec$”是：当 opt 表示 max 时表示“$<$”，当 opt 表示 min 时表示“$>$”。因而

$$\begin{aligned} V_{0,n-1}(s_0, p_{0,n-1}^*) &= V_{0,k-1}(s_0, p_{0,k-1}^*) + V_{k,n-1}(s_k^*, p_{k,n-1}^*) \\ &\prec V_{0,k-1}(s_0, p_{0,k-1}^*) + \operatorname*{opt}_{p_{k,n-1} \in P_{k,n-1}(s_k^*)} \{V_{k,n-1}(s_k^*, p_{k,n-1})\} \\ &\prec \operatorname*{opt}_{p_{0,k-1} \in P_{0,k-1}(s_0)} \{V_{0,k-1}(s_0, p_{0,k-1}) + \operatorname*{opt}_{p_{k,n-1} \in P_{k,n-1}(\tilde{s}_k)} V_{k,n-1}(\tilde{s}_k, p_{k,n-1})\} \end{aligned}$$

故与以上定理的必要性矛盾。证毕。

此推论就是前面提到的动态规划的“最优性原理”。它仅仅是最优策略的必要性。从最优性定理可以看到：如果一个决策问题有最优策略，则该问题的最优值函数一定可用动态规划的基本方程来表示，反之亦真。这是该定理为人们用动态规划方法去处理决策问题提供了理论依据和指明方法，就是要充分分析决策问题的结构，使它满足动态规划的条件，正确地写出动态规划的基本方程。

9.4 动态规划和静态规划的关系

动态规划、线性规划和非线性规划都是属于数学规划的范围，所研究的对象本质上都是一个求极值的问题，都是利用迭代法去逐步求解的。不过，线性规划和非线性规划所研究的问题，通常是与时间无关的，故又称它们为静态规划。线性规划迭代中的每一步是就问题的整体加以改善的。而动态规划所研究的问题是与时间有关的，它是研究具有多阶段决策过程的一类问题，将问题的整体按时间或空间的特征而分成若干个前后衔接的时空阶段，把多阶段决策问题表示为前后有关联的一系列单阶段决策问题，然后逐个加以解决，从而求出了整个问题的最优决策序列。因此，对于某些静态的问题，也可以人为地引入时间因素，把它看做是按阶段进行的一个动态规划问题，这就使得动态规划成为求解某些线性、非线性规划的有效方法。

由于动态规划方法有逆序解法和顺序解法之分，其关键在于正确写出动态规划的递推关系式，故递推方式有逆推和顺推两种形式。一般地说，当初始状态给定时，用逆推比较方便；当终止状态给定时，用顺推比较方便。

考查如图 9-6 所示的 n 阶段决策过程。

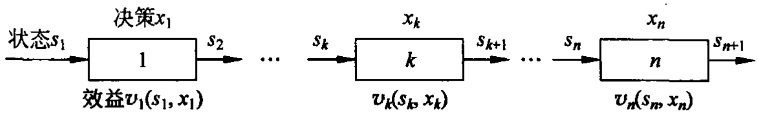


图 9-6

其中取状态变量为 $s_1,s_2,\cdots,s_{n+1}$；决策变量为 $x_1,x_2,\cdots,x_n$。在第 k 阶段，决策 x_k 使状态 s_k（输入）转移为状态 s_{k+1}（输出），设状态转移函数为

$$s_{k+1}=T_k(s_k,x_k),\quad k=1,2,\cdots,n$$

假定过程的总效益（指标函数）与各阶段效益（阶段指标函数）的关系为

$$V_{1,n}=v_1(s_1,x_1)*v_2(s_2,x_2)*\cdots*v_n(s_n,x_n)$$

其中记号“$*$”可都表示为“$+$”或者都表示为“$\times$”。问题为使 $V_{1,n}$ 达到最优化，即求 opt $V_{1,n}$，为简单起见，不妨此处就求 $\max V_{1,n}$。

9.4.1 逆推解法

设已知初始状态为 s_1，并假定最优值函数 $f_k(s_k)$ 表示第 k 阶段的初始状态为 s_k，从 k 阶段到 n 阶段所得到的最大效益。

从第 n 阶段开始，则有

$$f_n(s_n)=\max_{x_n\in D_n(s_n)} v_n(s_n,x_n)$$

其中 $D_n(s_n)$ 是由状态 s_n 所确定的第 n 阶段的允许决策集合。解此一维极值问题，就得到最优解 $x_n=x_n(s_n)$ 和最优值 $f_n(s_n)$，要注意的是，若 $D_n(s_n)$ 只有一个决策，则 $x_n\in D_n(s_n)$ 就应写成 $x_n=x_n(s_n)$。

在第 $n-1$ 阶段，有

$$f_{n-1}(s_{n-1})=\max_{x_{n-1}\in D_n(s_{n-1})}[v_{n-1}(s_{n-1},x_{n-1})*f_n(s_n)]$$

其中 $s_n=T_{n-1}(s_{n-1},x_{n-1})$；解此一维极值问题，得到最优解 $x_{n-1}=x_{n-1}(s_{n-1})$ 和最优值 $f_{n-1}(s_{n-1})$。

在第 k 阶段，有

$$f_k(s_k)=\max_{x_k\in D_k(s_k)}[v_k(s_k,x_k)*f_{k+1}(s_{k+1})]$$

其中 $s_{k+1}=T_k(s_k,x_k)$。解得最优解 $x_k=x_k(s_k)$ 和最优值 $f_k(s_k)$。

如此类推，直到第一阶段，有

$$f_1(s_1)=\max_{x_1\in D_1(s_1)}[v_1(s_1,x_1)*f_2(s_2)]$$

其中 $s_2=T_1(s_1,x_1)$；解得最优解 $x_1=x_1(s_1)$ 和最优值 $f_1(s_1)$。

由于初始状态 s_1 已知，故 $x_1=x_1(s_1)$ 和 $f_1(s_1)$ 是确定的，从而 $s_2=T_1(s_1,x_1)$ 也就可确定，于是 $x_2=x_2(s_2)$ 和 $f_2(s_2)$ 也就可确定。这样，按照上述递推过程相反的顺序推算下

去，就可逐步确定出每阶段的决策及效益。

例 3 用逆推解法求解下面问题

$$\max z = x_1 x_2^2 x_3$$

$$\begin{cases} x_1 + x_2 + x_3 = c \quad (c>0) \\ x_i \geqslant 0, \quad i=1,2,3 \end{cases}$$

解 按问题的变量个数划分阶段，把它看做为一个三阶段决策问题。设状态变量为 s_1、s_2、s_3、s_4，并记 $s_1=c$；取问题中的变量 x_1、x_2、x_3 为决策变量；各阶段指标函数按乘积方式结合。令最优值函数 $f_k(s_k)$ 表示为第 k 阶段的初始状态为 s_k，从 k 阶段到 3 阶段所得到的最大值。

设 $\quad s_3 = x_3 \qquad s_3 + x_2 = s_2 \qquad s_2 + x_1 = s_1 = c$

则有 $\quad x_3 = s_3 \qquad 0 \leqslant x_2 \leqslant s_2 \qquad 0 \leqslant x_1 \leqslant s_1 = c$

于是用逆推解法，从后向前依次有

$$f_3(s_3) = \max_{x_3 = s_3}(x_3) = s_3 \text{ 及最优解 } x_3^* = s_3$$

$$f_2(s_2) = \max_{0 \leqslant x_2 \leqslant s_2}[x_2^2 \cdot f_3(s_3)] = \max_{0 \leqslant x_2 \leqslant s_2}[x_2^2(s_2 - x_2)] = \max_{0 \leqslant x_2 \leqslant s_2} h_2(s_2, x_2)$$

由 $\dfrac{\mathrm{d}h_2}{\mathrm{d}x_2} = 2x_2 s_2 - 3x_2^2 = 0$，得 $x_2 = \dfrac{2}{3}s_2$ 和 $x_2 = 0$（舍去）

又 $\dfrac{\mathrm{d}^2 h_2}{\mathrm{d}x_2^2} = 2s_2 - 6x_2$，而 $\left.\dfrac{\mathrm{d}^2 h_2}{\mathrm{d}x_2^2}\right|_{x_2=\frac{2}{3}s_2} = -2s_2 < 0$，故 $x_2 = \dfrac{2}{3}s_2$ 为极大值点。

所以 $f_2(s_2) = \dfrac{4}{27}s_2^3$ 及最优解 $x_2^* = \dfrac{2}{3}s_2$

$$f_1(s_1) = \max_{0 \leqslant x_1 \leqslant s_1}[x_1 \cdot f_2(s_2)] = \max_{0 \leqslant x_1 \leqslant s_1}\left[x_1 \cdot \frac{4}{27}(s_1 - x_1)^3\right]$$
$$= \max_{0 \leqslant x_1 \leqslant s_1} h_1(s_1, x_1)$$

像前面一样利用微分法易知 $\qquad x_1^* = \dfrac{1}{4}s_1$

故 $\qquad f_1(s_1) = \dfrac{1}{64}s_1^4$

由于已知 $s_1 = c$，因而按计算的顺序反推算，可得各阶段的最优决策和最优值。即

$$x_1^* = \frac{1}{4}c, \quad f_1(c) = \frac{1}{64}c^4$$

由

$$s_2 = s_1 - x_1^* = c - \frac{1}{4}c = \frac{3}{4}c$$

所以

$$x_2^* = \frac{2}{3}s_2 = \frac{1}{2}c, \quad f_2(s_2) = \frac{1}{16}c^3$$

由

$$s_3=s_2-x_2^*=\frac{3}{4}c-\frac{1}{2}c=\frac{1}{4}c$$

所以

$$x_3^*=\frac{1}{4}c,\quad f_3(s_3)=\frac{1}{4}c$$

因此得到最优解为 $x_1^*=\frac{1}{4}c,\quad x_2^*=\frac{1}{2}c,\quad x_3^*=\frac{1}{4}c$

最大值为 $\max z=f_1(c)=\frac{1}{64}c^4$

9.4.2 顺推解法

设已知终止状态 s_{n+1}，并假定最优值函数 $f_k(s)$ 表示第 k 阶段末的结束状态为 s，从 1 阶段到 k 阶段所得到的最大收益。

已知终止状态 s_{n+1} 用顺推解法与已知初始状态用逆推解法在本质上没有区别，它相当于把实际的起点视为终点，实际的终点视为起点，而按逆推解法进行的。换言之，只要把图 9-6 的箭头倒转过来即可，把输出 s_{k+1} 看做输入，把输入 s_k 看做输出，这样便得到顺推解法。但应注意，这里是在上述状态变量和决策变量的记法不变的情况下考虑的。因而这时的状态变换是上面状态变换的逆变换，记为 $s_k=T_k^*(s_{k+1},x_k)$；从运算而言，即是由 s_{k+1} 和 x_k 而去确定 s_k 的。

从第一阶段开始，有

$$f_1(s_2)=\max_{x_1\in D_1(s_2)} v_1(s_2,x_1),\quad \text{其中 } s_1=T_1^*(s_2,x_1)$$

解得最优解 $x_1=x_1(s_2)$ 和最优值 $f_1(s_2)$。若 $D_1(s_2)$ 只有一个决策，则 $x_1\in D_1(s_2)$ 就写成 $x_1=x_1(s_2)$。

在第二阶段，有

$$f_2(s_3)=\max_{x_2\in D_2(s_3)}[v_2(s_3,x_2)*f_1(s_2)]$$

式中 $s_2=T_2^*(s_3,x_2)$，解得最优解 $x_2=x_2(s_3)$ 和最优值 $f_2(x_3)$。

如此类推，直到第 n 阶段，有

$$f_n(s_{n+1})=\max_{x_n\in D_n(s_{n+1})}[v_n(s_{n+1},x_n)*f_{n-1}(s_n)]$$

其中 $s_n=T_n^*(s_{n+1},x_n)$。解得最优解 $x_n=x_n(s_{n+1})$ 和最优值 $f_n(s_{n+1})$。

由于终止状态 s_{n+1} 是已知的，故 $x_n=x_n(s_{n+1})$ 和 $f_n(s_{n+1})$ 是确定的。再按计算过程的相反顺序推算上去，就可逐步确定出每阶段的决策及效益。

应指出的是，若将状态变量的记法改为 $s_0,s_1,\cdots,s_n$，决策变量记法不变，则按顺序解法，此时的最优值函数为 $f_k(s_k)$。因而，这个符号与逆推解法的符号一样，但含义是不同的，这里的 s_k 是表示 k 阶段末的结束状态。

例 4 将例 3 用顺推解法解之。

解 设 $s_4=c$，令最优值函数 $f_k(s_{k+1})$ 表示第 k 阶段末的结束状态为 s_{k+1}，从 1 阶段到 k 阶段的最大值。

设 $$s_2=x_1,s_2+x_2=s_3,s_3+x_3=s_4=c$$

则有 $$x_1=s_2,0\leqslant x_2\leqslant s_3,0\leqslant x_3\leqslant s_4$$

于是用顺推解法，从前向后依次有

$$f_1(s_2)=\max_{x_1=s_2}(x_1)=s_2 \text{ 及最优解 } x_1^*=s_2。$$

$$f_2(s_3)=\max_{0\leqslant x_2\leqslant s_3}[x_2^2 f_1(s_2)]=\max_{0\leqslant x_2\leqslant s_3}[x_2^2(s_3-x_2)]=\frac{4}{27}s_3^3$$

及最优解 $x_2^*=\frac{2}{3}s_3$。

$$f_3(s_4)=\max_{0\leqslant x_3\leqslant s_4}[x_3 f_2(s_3)]=\max_{0\leqslant x_3\leqslant s_4}\left[x_3\ \frac{4}{27}(s_4-x_3)^3\right]=\frac{1}{64}s_4^4$$

及最优解 $x_3^*=\frac{1}{4}s_3$。

由于已知 $s_4=c$，故易得到最优解为 $x_1^*=\frac{1}{4}c,x_2^*=\frac{1}{2}c,x_3^*=\frac{1}{4}c$；相应的最大值为

$$\max z=\frac{1}{64}c^4$$

现在再考虑若是已知初始状态 $s_1=c$，将例 3 用顺推解法又如何进行呢？

因这时的状态转移函数为 $s_k=s_{k+1}+x_k,k=1,2,3$。为了保证决策变量非负，必须有 $s_{k+1}\leqslant s_k\leqslant c$。

因此，设 $$x_1+s_2=s_1=c,x_2+s_3=s_2,x_3+s_4=s_3$$

则有 $$x_1=s_1-s_2=c-s_2,0\leqslant x_2=s_2-s_3\leqslant c-s_3,$$
$$0\leqslant x_3=s_3-s_4\leqslant c-s_4$$

于是用顺推解法，从前向后依次有

$$f_1(s_2)=\max_{x_1=c-s_2}(x_1)=c-s_2 \quad \text{及最优解} \quad x_1^*=c-s_2$$

$$f_2(s_3)=\max_{0\leqslant x_2\leqslant c-s_3}[x_2^2 f_1(s_2)]=\max_{0\leqslant x_2\leqslant c-s_3}[x_2^2(c-s_3-x_2)]$$
$$=\frac{4}{27}(c-s_3)^3 \quad \text{及最优解} \quad x_2^*=\frac{2}{3}(c-s_3)$$

$$f_3(s_4)=\max_{0\leqslant x_3\leqslant c-s_4}[x_3 f_2(s_3)]=\max_{0\leqslant x_3\leqslant c-s_4}\left[x_3\cdot\frac{4}{27}(c-s_4-x_3)^3\right]$$
$$=\frac{1}{64}(c-s_4)^4 \quad \text{及最优解} \quad x_3^*=\frac{1}{4}(c-s_4)$$

由于终止状态 s_4 不知道，故须再对 s_4 求一次极值，即

$$\max_{0\leqslant s_4\leqslant c} f_3(s_4)=\max_{0\leqslant s_4\leqslant c}\frac{1}{64}(c-s_4)^4$$

显然,只有当 $s_4=0$ 时,$f_3(s_4)$才能达到最大值。然后按计算顺序反推算可求出各阶段的最优决策及最优值。最后得到最优解为 $x_1^*=\frac{1}{4}c,x_2^*=\frac{1}{2}c,x_3^*=\frac{1}{4}c$;最大值为

$$\max z=f_3(0)=\frac{1}{64}c^4$$

注意 若记状态变量为 s_0、s_1、s_2、s_3,取 $s_0=c$;决策变量记法不变;令最优值函数 $f_k(s_k)$表示第 k 阶段末的结束状态为 s_k,从 1 阶段到 k 阶段的最大值,则按顺推解法,从前向后依次为

$$f_1(s_1)=\max_{x_1=c-s_1}(x_1)$$

$$f_2(s_2)=\max_{0\leqslant x_2\leqslant c-s_2}[x_2^2f_1(s_1)]$$

$$f_3(s_3)=\max_{0\leqslant x_3\leqslant c-s_3}[x_3f_2(s_2)]$$

例 5 用动态规划方法解下面问题

$$\max F=4x_1^2-x_2^2+2x_3^2+12$$

$$\begin{cases}3x_1+2x_2+x_3\leqslant 9\\ x_i\geqslant 0,\quad i=1,2,3\end{cases}$$

解 按问题中变量的个数分为三个阶段。设状态变量为 s_0、s_1、s_2、s_3,并记 $s_3\leqslant 9$;取 x_1、x_2、x_3 为各阶段的决策变量;各阶段指标函数按加法方式结合。令最优值函数 $f_k(s_k)$表示第 k 阶段的结束状态为 s_k,从 1 阶段至 k 阶段的最大值。

设 $$3x_1=s_1,s_1+2x_2=s_2,s_2+x_3=s_3\leqslant 9$$

则有 $$x_1=s_1/3,0\leqslant x_2\leqslant s_2/2,0\leqslant x_3\leqslant s_3$$

于是用顺推方法,从前向后依次有

$$f_1(s_1)=\max_{x_1=s_1/3}(4x_1^2)=\frac{4}{9}s_1^2\quad\text{及最优解}\quad x_1^*=\frac{s_1}{3}$$

$$f_2(s_2)=\max_{0\leqslant x_2\leqslant s_2/2}[-x_2^2+f_1(s_1)]=\max_{0\leqslant x_2\leqslant s_2/2}\left[-x_2^2+\frac{4}{9}(s_2-2x_2)^2\right]$$

$$=\max_{0\leqslant x_2\leqslant s_2/2}h_2(s_2,x_2)$$

由 $$\frac{\mathrm{d}h_2}{\mathrm{d}x_2}=\frac{14}{9}x_2-\frac{16}{9}s_2=0\quad\text{解得}\quad x_2=\frac{8}{7}s_2$$

因该点不在允许决策集合内,故无须判别。因而 $h_2(s_2,x_2)$的最大值必在两个端点上选取。

而 $$h_2(0)=\frac{4}{9}s_2^2,h_2\left(\frac{s_2}{2}\right)=-\frac{s_2^2}{4}$$

所以 $h_2(s_2,x_2)$的最大值点在 $x_2=0$ 处,故得到 $f_2(s_2)=\frac{4}{9}s_2^2$ 及相应的最优解$x_2^*=0$。

$$f_3(s_3)=\max_{0\leqslant x_3\leqslant s_3}[2x_3^2+12+f_2(s_2)]$$

$$=\max_{0\leqslant x_3\leqslant s_3}\left[2x_3^2+12+\frac{4}{9}(s_3-x_3)^2\right]=\max_{0\leqslant x_3\leqslant s_3}h_3(s_3,x_3)$$

由
$$\frac{\mathrm{d}h_3}{\mathrm{d}x_3}=\frac{44}{9}x_3-\frac{8}{9}s_3=0\quad 解得\quad x_3=\frac{2}{11}s_3$$

又
$$\frac{\mathrm{d}^2h_3}{\mathrm{d}x_3^2}=\frac{44}{9}>0，故该点为极小值点。$$

而
$$h_3(0)=\frac{4}{9}s_3^2+12,\quad h_3(s_3)=2s_3^2+12$$

故 $h_3(s_2,x_3)$ 的最大值点在 $x_3=s_3$ 处，所以得 $f_3(x_3)=2s_3^2+12$ 及相应的最优解 $x_3^*=s_3$。

由于 s_3 不知道，故须再对 s_3 求一次极值，即

$$\max_{0\leqslant s_3\leqslant 9}f_3(s_3)=\max_{0\leqslant s_3\leqslant 9}[2s_3^2+12]$$

显然，当 $s_3=9$ 时 $f_3(s_3)$ 才能达到最大值。所以 $f_1(9)=2\times 9^2+12=174$ 为最大值。

再按计算的顺序反推算可求得最优解为 $x_1^*=0,x_2^*=0,x_3=9$；最大值为

$$\max F=f_1(9)=174$$

注意

(1) 若先作代换，令 $y_1=3x_1,y_2=2x_2,y_3=x_3$，将原问题变为

$$\max F=\frac{4}{9}y_1^2-\frac{1}{4}y_2^2+2y_3^2+12$$

$$\begin{cases}y_1+y_2+y_3\leqslant 9\\ y_i\geqslant 0,\quad i=1,2,3\end{cases}$$

再解此问题，然后换回 x_i 也可以。

(2) 在计算 h_2 和 h_3 的最大值中，若学习了凸函数的性质，也可利用凸函数的性质来确定最大值。

9.4.3　动态规划的计算框图

在实际问题中，函数序列 $f_k(s_k)$ 往往不能表示为解析形式；状态变量 s_k 和决策变量 x_k 即使是离散的，其集合也很大。这样，求 $f_k(s_k)$ 和最优策略的数值解计算量就很大，一般要用计算机来解决。

下面讨论逆序解法的迭代计算程序。由于始端可能是自由状态或固定状态两种情况，终端也可能是自由状态或固定状态两种情况，因而组合起来就有四种情况。这里，仅讨论终端自由而且始端自由或固定的两种情况的计算框图。至于终端固定的情况框图类似，就不讨论了。

为明确起见，假设：决策变量 x_k 是连续实变数，允许决策集合 $D_k(s_k)$ 是实数集合；状态变量 s_k 也是连续实变数，各段可达状态集合 S_k 的实数闭区间是 $[\underline{s}_k,\bar{s}_k]$，对于始端固定的状态变量，各段可达状态集合 s_k^1 的实数闭区间是 $[\underline{s}_k^1,\bar{s}_k^1]$。在计算前，先要把 s_k 离散

化。为此，要选好适当的增量 Δs。在第 k 段，计算将在点列 $\{\underline{s}_k, \underline{s}_k+\Delta s, \cdots, \underline{s}_k+m_k\Delta s\}$ 上进行，其中 m_k 是满足

$$\underline{s}_k+m_k\Delta s\leqslant \bar{S}_k\leqslant \bar{S}_k+(m_k+1)\Delta s \text{ 的正整数。}$$

根据一般的逆序动态规划的基本方程：

$$\begin{cases} f_{n+1}(s_{n+1})=0 \\ f_k(s_k)=\underset{x_k\in D_k(s_k)}{\text{opt}}\{v_k(s_k,x_k)+f_{k+1}(s_{k+1})\}, \quad k=n,\cdots,2,1 \end{cases}$$

其中 $s_{k+1}=T_k(s_k,s_k)$，且当 s_1 固定时，$s_k\in S_k^1$；当 s_1 自由时，$s_k\in S_k$。因而可得动态规划的逆序解法计算程序框图(见图 9-7)，下面对框图作几点说明。

(1) 图 9-7 中，左边部分是指标函数序列$\{f_k(s_k)\}$的递推计算，它是逆序的，即 k 由 n 逐步减少到 1；右边部分是最优决策序列$\{x_k^*\}$的递推计算，它是顺序的，即 k 由 1 逐步增大到 n。

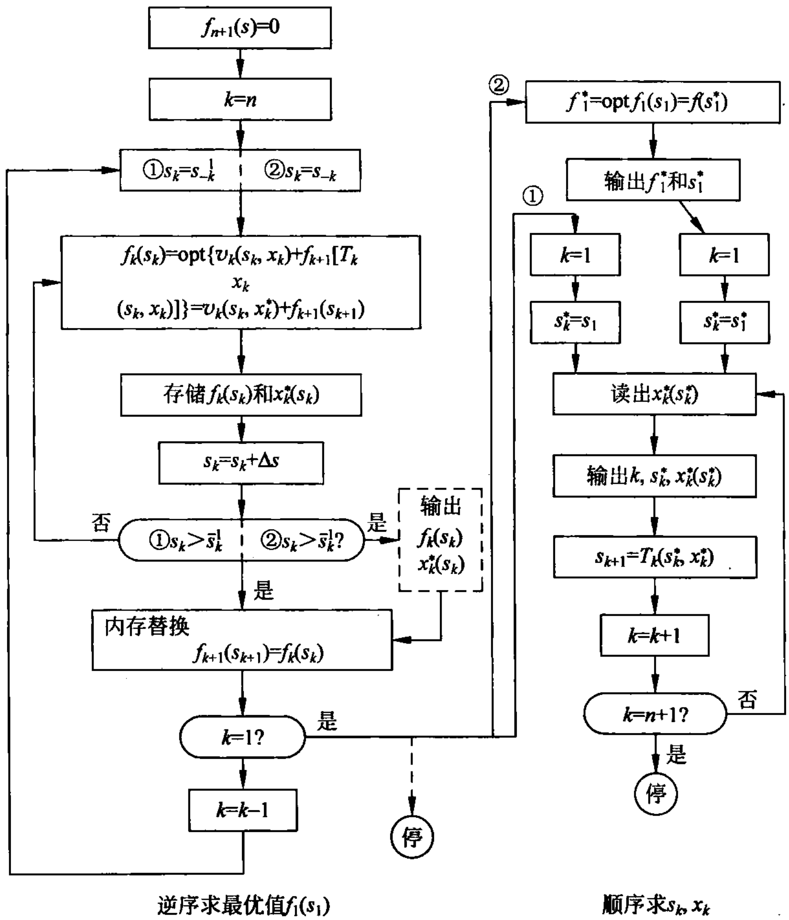


图 9-7 动态规划逆序解法计算程序框图

(2) 若实际问题要求的不只是整个过程的最优解，而且要求出从各段出发的最优策略和最优值。则在程序中，对每一 k 段，计算完 $f_k(s_k)$ 和 $x_k^*(s_k)$ 后就可输出，如图中的虚线所示。如不需要输出 s_k，可把右边部分框图取消。

(3) 框图中包含①固定始端和②自由始端两种情况。它们的区别是，左边部分输入数据不同，右边部分在自由始端②情况下，还需多求一次最优值计算。

(4) 因 $f_{k+1}(s_{k+1})$ 只在计算 k 段时有用，到 $k-1$ 段就没用了。故在计算 k 段时，$f_k(s_k)$ 都要存入内存，在计算 $k-1$ 段时，可用 $f_k(s_k)$ 把 $f_{k+1}(s_{k+1})$ 替换掉。函数 $x_k^*(s_k)$ 在左边部分计算出来后不要用，可送入外存。在右边部分求 $\{x_k^*\}$ 需用时，再依 k 的序列将 x_k 由外存移入内存。

(5) 在计算 $f_k(s_k)$ 时，s_k 在点列上取值，对于

$$点\ s_k=\underline{s}_k+j\Delta s(0\leqslant j\leqslant m_k)$$
$$点\ \tilde{s}_{k+1}=T_k(s_k,x_k)(x_k\in D_k(s_k))$$

不一定在 s_{k+1} 的点列中，这时，必须选择适当的内插公式，由 $f_{k+1}(s_{k+1})$ 在点列上的值求它在点 $\tilde{s}_{k+1}$ 上的值。

逆序解法计算程序框图：①固定始端；②自由始端。

最后应指出的是，在这节里运用递推关系逐步求出极值函数 $f_1(s)$、…、$f_n(s)$ 及相应的决策函数 $x_1(s)$、…、$x_n(s)$，这是一种通过函数值不断的迭代过程，而逐步达到最优值，通常称为**函数空间迭代法**。这种迭代方法，不仅对像例 1 那样阶段数为确定有限值的**定期多阶段决策过程**有效，而且，对在实际问题中，出现的动态规划基本方程，不是一个递推方程，而是为某函数的泛函方程，那种阶段数为有限但不固定的**不定期多阶段决策过程**或阶段数为无限(或很大)的无期多阶段决策过程也是一种重要的求解方法。还有对解上述两类过程比函数空间迭代法的收敛速度要快些的**策略空间迭代法**。函数空间迭代法和策略空间迭代法都是动态规划求解不定期或无期的多阶段决策过程的两种重要方法。由于篇幅有限，这里就不介绍了。

注　记

通过这章的学习，大家看到：任何一个多阶段决策过程的最优化问题，都可以用非线性规划(特殊的为线性规划)模型来描述。从原则上说，一般也可以用非线性规划方法来求解。那么，用动态规划方法有什么优越性呢？

(1) 易于确定全局最优解。即使指标函数形式较简单，由于约束条件所确定的约束集合往往十分复杂，故应用目前的非线性规划方法求全局最优解是非常困难的。而动态规划方法是一种逐步改善法，它把原问题化成一系列结构相似的最优化子问题，而每个子问题的变量个数比原问题少得多，约束集合也简单得多，故较易于确定全局最优解。特别是，对于一类其指标、状态转移和允许决策集合不能用分析形式表出的过程最优化问题(如非线性整数规划、离散模型)，用分析方法无法求出最优解，而用动态规划却很容易。基于这点，目前有相当多的最优化问题，动态规划是求出其全局最优解的唯一方法。

(2) 能得到一族解，有利于分析结果。非线性规划的方法是对问题的整体求解，是单阶段进行的，它只能得到全过程的解。而动态规划方法是将求解分成多阶段进行，求出的不仅是全过程的解，而且包

括所有子过程的一族解。在某些情况下，这些解族正是实际问题所需要的，它有助于分析结果是否有用等，这时，动态规划方法比其他方法更显示出优越性，且大大节省了计算量。

(3) 能利用经验，提高求解的效率。动态规划方法反映了过程逐段演变的前后联系，较之非线性规划与实际过程联系得更紧密，因而在计算中，能更有效地利用经验，提高求解的效率。如在策略空间迭代法中，初始策略的选取对于迭代的收敛速度有很大的影响，故利用经验能帮助选好初始策略。有的问题，不一定要知道最优解的值，而关心的是最优解的结构或过程中某些参数的依赖关系等，这时利用动态规划方法分析较方便。

动态规划方法也存在不足之处：

(1) 到目前为止，没有一个统一的标准模型可供应用。由于实际问题不同，其动态规划模型也就有异，虽然理论上说可以把某些静态规划的问题转化为动态规划模型来求解，但这种转化有时将变得十分困难，需要丰富的想象力和灵活的技巧性。

(2) 应用的局限性。由于构造动态规划模型时，状态变量必须满足"无后效性"条件，这条件不仅依赖于状态转移规律，还依赖于允许决策集合和指标函数的结构，是一个相当强的条件。不少实际问题在取其自然特征作为状态变量往往不能满足这条件，这就降低了动态规划的通用性。虽然，从原则上说，采取适当增多状态变量的办法，总能人为地把过程变为无后效的，但由于动态规划的常规算法还存在局限性，故这种做法目前不可取。

(3) 在数值求解时，存在"维数障碍"。由于在大多数问题中，函数$\{f_k(s_k)\}$必须以状态集合s_k内格点上的数值来表示。故在电子计算机上计算时，每递推一段，都必须把前一段算出的最优值函数在相应的状态集合上的全部值存入内存中。设每一状态变量的集合是实数区间，如果每区间都分成$m-1$个格子，则在一维情况下，表示f_k之值的格点数为m，而在n维情况下，格点数为m^n。可见当维数增大时，所需的内存量成指数倍增长。例如，若计算机的内存量为$2^{16}=32\ 768$字节，由

维数 n	1	2	3	4	5
每维格点数	32 768	181	32	13	8

可见，在这样内存限制下，超过三维(包括三维)的问题用通常的动态规划方法是不可取的。为了克服"维数障碍"，目前已找出了不少方法。如多项式逼近法、逐次逼近法、状态微增法、拉格朗日乘数法等等。它们对于不同范围内的问题已经证明在一定程度上可以克服这个障碍。但是目前尚无一般的解决办法。

应指出的是，动态规划方法虽存在许多不足之处，但应用是广泛的。许多专家学者结合实际已成功解决了许多实际问题。我国情况也是如此，如张勇传院士等，在1985年前，将动态规划方法应用于水电站水库优化调度方面作出了成绩，曾获1985年国家科委颁布的一等奖(见《光明日报》1985年10月8日)。

随着计算机的日益普及，用动态规划基本思想去处理问题更加广泛。许多介绍算法的书里，都提到使用动态规划方法，如何巧妙地解决了科学技术和实际生活中问题的例子，称这种算法是高效率的算法之一。

习题

9.1 试分析本章第9.1节之例2问题中(1)阶段的划分；(2)状态变量和它的取值范围；(3)决策变量和它的允许决策集合；(4)状态转移方程；(5)指标函数和最优值函数。

9.2 设某工厂自国外进口一部精密机器，由机器制造厂至出口港有三个港口可供选择，而进口港又有三个可供选择，进口后可经由两个城市到达目的地，其间的运输成本如图 9-8 中所标的数字，试求运费最低的路线？

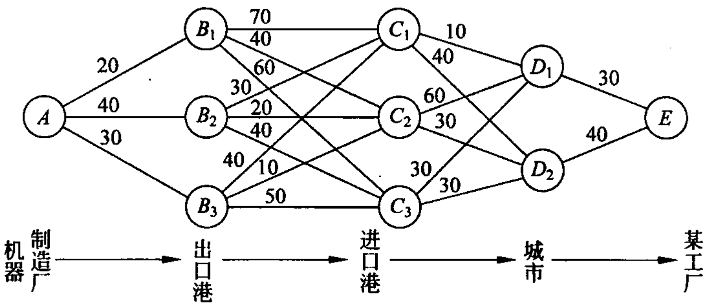


图 9-8

9.3 计算从 A 到 B、C 和 D 的最短路线。已知各段路线的长度如图 9-9 所示。

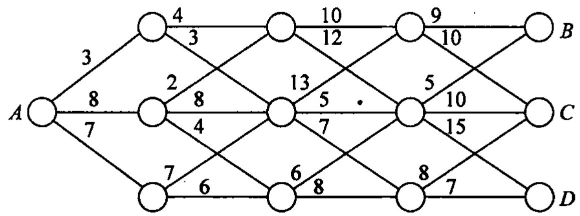


图 9-9

9.4 写出下列问题的动态规划的基本方程。

(1) $\max z=\sum_{i=1}^{n}\phi_i(x_i)$

$$\begin{cases}\sum_{i=1}^{n}x_i=b \quad (b>0)\\ x_i\geqslant 0,\quad (i=1,2,\cdots,n)\end{cases}$$

(2) $\min z=\sum_{i=1}^{n}c_ix_i^2$

$$\begin{cases}\sum_{i=1}^{n}a_ix_i\geqslant b \quad (a_i>0)\\ x_i\geqslant 0,\quad (i=1,2,\cdots,n)\end{cases}$$

9.5 用递推方法求解下列问题。

(1) $\max z=4x_1+9x_2+2x_3^2$

$$\begin{cases}x_1+x_2+x_3=10\\ x_i\geqslant 0,\quad i=1,2,3\end{cases}$$

(2) $\max z=4x_1+9x_2+2x_3^2$

$$\begin{cases}2x_1+4x_2+3x_3\leqslant 10\\ x_i\geqslant 0,\quad i=1,2,3\end{cases}$$

(3) $\max z=x_1\cdots x_n$

$$\begin{cases}\sum_{i=1}^{n}x_i=c \quad (c>0)\\ x_i\geqslant 0,\quad i=1,2,\cdots,n\end{cases}$$

(4) $\min z=3x_1-5x_1+3x_2^2-3x_2+2x_3^2-7x_3$

$$\begin{cases}2x_1+3x_2+2x_3\geqslant 16\\ x_i\geqslant 0,\quad i=1,2,3\end{cases}$$

(5) $\max z=3x_1^3-4x_1+2x_2^2-5x_2+2x_3$

$$\begin{cases}4x_1+2x_2+3x_3\leqslant 18\\ x_i\geqslant 0,\quad i=1,2,3\end{cases}$$

(6) $\min z=\sum_{i=1}^{n} x_i^p\ (p>1)$

$$\begin{cases}\sum_{i=1}^{n} x_i=c\quad (c>0)\\ x_i\geqslant 0,\quad i=1,2,\cdots,n\end{cases}$$

9.6 设某人有 400 万元金额,计划在四年内全部用于投资中去。已知在一年内若投资用去 x 万元就能获得$\sqrt{x}$万元的效用。每年没有用掉的金额,连同利息(年利息 10%)可再用于下一年的投资。而每年已打算用于投资的金额不计利息。试制定金额的使用计划,而使四年内获得的总效用最大?

(1) 用动态规划方法求解;

(2) 用拉格朗日乘数法求解;

(3) 比较两种解法,并说明动态规划方法有哪些优点。

CHAPTER 10

第10章 动态规划应用举例

10.1 资源分配问题

所谓分配问题，就是将数量一定的一种或若干种资源（例如原材料、资金、机器设备、劳力、食品等等），恰当地分配给若干个使用者，而使目标函数为最优。

10.1.1 资源分配问题

设有某种原料，总数量为 a，用于生产 n 种产品。若分配数量 x_i 用于生产第 i 种产品，其收益为 $g_i(x_i)$。问应如何分配，才能使生产 n 种产品的总收入最大？

此问题可写成静态规划问题

$$\begin{cases}\max z=g_1(x_1)+g_2(x_2)+\cdots+g_n(x_n)\\ x_1+x_2+\cdots+x_n=a\\ x_i\geqslant 0,\quad i=1,2,\cdots,n\end{cases}$$

当 $g_i(x_i)$ 都是线性函数时，它是一个线性规划问题；当 $g_i(x_i)$ 是非线性函数时，它是一个非线性规划问题。但当 n 比较大时，具体求解是比较麻烦的。然而，由于这类问题的特殊结构，可以将它看成一个多阶段决策问题，并利用动态规划的递推关系来求解。

在应用动态规划方法处理这类"静态规划"问题时，通常以把资源分配给一个或几个使用者的过程作为一个阶段，把问题中的变量 x_i 选为决策变量，将累计的量或随递推过程变化的量选为状态变量。

设状态变量 s_k 表示分配用于生产第 k 种产品至第 n 种产品的原料数量。

决策变量 u_k 表示分配给生产第 k 种产品的原料数，即 $u_k=x_k$

状态转移方程

$$s_{k+1}=s_k-u_k=s_k-x_k$$

允许决策集合

$$D_k(s_k)=\{u_k\,|\,0\leqslant u_k=x_k\leqslant s_k\}$$

令最优值函数 $f_k(s_k)$ 表示以数量为 s_k 的原料分配给第 k 种产品至第 n 种产品所得到的最大总收入。因而可写出动态规划的逆推关系式为

$$\begin{cases} f_k(s_k)=\max\limits_{0\leqslant x_k\leqslant s_k}\{g_k(x_k)+f_{k+1}(s_k-x_k)\}, & k=n-1,\cdots,1 \\ f_n(s_n)=\max\limits_{x_n=s_n} g_n(x_n) \end{cases}$$

利用这个递推关系式进行逐段计算,最后求得 $f_1(a)$ 即为所求问题的最大总收入。

例 1 某工业部门根据国家计划的安排,拟将某种高效率的设备五台,分配给所属的甲、乙、丙三个工厂,各工厂若获得这种设备之后,可以为国家提供的赢利如表 10-1 所示。

表 10-1 某工业部门设备高效利用赢利表 万元

设备台数	工厂		
	甲	乙	丙
0	0	0	0
1	3	5	4
2	7	10	6
3	9	11	11
4	12	11	12
5	13	11	12

问:这五台设备如何分配给各工厂,才能使国家得到的赢利最大。

解 将问题按工厂分为三个阶段,甲、乙、丙三个工厂分别编号为 1、2、3

设 s_k 表示为分配给第 k 个工厂至第 n 个工厂的设备台数。

x_k 表示为分配给第 k 个工厂的设备台数。

则 $s_{k+1}=s_k-x_k$ 为分配给第 $k+1$ 个工厂至第 n 个工厂的设备台数。

$P_k(x_k)$表示为 x_k 台设备分配到第 k 个工厂所得的赢利值。

$f_k(s_k)$表示为 s_k 台设备分配给第 k 个工厂至第 n 个工厂时所得到的最大赢利值。

因而可写出逆推关系式为

$$\begin{cases} f_k(s_k)=\max\limits_{0\leqslant x_k\leqslant s_k}[P_k(x_k)+f_{k+1}(s_k-x_k)], & k=3,2,1 \\ f_4(s_4)=0 \end{cases}$$

下面从最后一个阶段开始向前逆推计算。

第三阶段:

设将 s_3 台设备($s_3=0,1,2,3,4,5$)全部分配给工厂丙时,则最大赢利值为

$$f_3(s_3)=\max_{x_3}[P_3(x_3)]$$

其中 $x_3=s_3=0,1,2,3,4,5$

因为此时只有一个工厂,有多少台设备就全部分配给工厂丙,故它的赢利值就是该段的最大赢利值。其数值计算如表 10-2 所示。

表　10-2

s_3	$P_3(x_3)$						$f_3(s_3)$	x_3^*
	x_3							
	0	1	2	3	4	5		
0	0						0	0
1		4					4	1
2			6				6	2
3				11			11	3
4					12		12	4
5						12	12	5

表中 x_3^* 表示使 $f_3(s_3)$为最大值时的最优决策。

第二阶段：

设把 s_2 台设备(s_2=0,1,2,3,4,5)分配给工厂乙和工厂丙时，则对每个 s_2 值，有一种最优分配方案，使最大赢利值为

$$f_2(s_2)=\max_{x_2}[P_2(x_2)+f_3(s_2-x_2)]$$

其中　　$x_2=0,1,2,3,4,5$

因为给乙工厂 x_2 台，其赢利为 $P_2(x_2)$，余下的 s_2-x_2 台就给丙工厂，则它的赢利最大值为 $f_3(s_2-x_2)$。现要选择 x_2 的值，使 $P_2(x_2)+f_3(s_2-x_2)$取最大值。其数值计算如表 10-3所示。

表　10-3

s_2	$P_2(x_2)+f_3(s_2-x_2)$						$f_2(s_2)$	x_2^*
	x_2							
	0	1	2	3	4	5		
0	0						0	0
1	0+4	5+0					5	1
2	0+6	5+4	10+0				10	2
3	0+11	5+6	10+4	11+0			14	2
4	0+12	5+11	10+6	11+4	11+0		16	1,2
5	0+12	5+12	10+11	11+6	11+4	11+0	21	2

第一阶段：

设把 s_1 台(这里只有 $s_1=5$ 的情况)设备分配给甲、乙、丙三个工厂时，则最大赢

利值为

$$f_1(5)=\max_{x_1}[P_1(x_1)+f_2(5-x_1)]$$

其中 $$x_1=0,1,2,3,4,5$$

因为给甲工厂 x_1 台，其赢利为 $P_1(x_1)$，剩下的 $5-x_1$ 台就分给乙和丙两个工厂，则它的赢利最大值为 $f_2(5-x_1)$。现要选择 x_1 值，使 $P_1(x_1)+f_2(5-x_1)$ 取最大值，它就是所求的总赢利最大值，其数值计算如表 10-4 所示。

表 10-4

s_1	$P_1(x_1)+f_2(5-x_1)$						$f_1(5)$	x_1^*
	x_1							
	0	1	2	3	4	5		
5	0+21	3+16	7+14	9+10	12+5	13+0	21	0,2

然后按计算表格的顺序反推算，可知最优分配方案有两个：

(1) 由于 $x_1^*=0$，根据 $s_2=s_1-x_1^*=5-0=5$，查表 10-3 知 $x_2^*=2$，由 $s_3=s_2-x_2^*=5-2=3$，故 $x_3^*=s_3=3$。即得甲工厂分配 0 台，乙工厂分配 2 台，丙工厂分配3 台。

(2) 由于 $x_1^*=2$，根据 $s_2=s_1-x_1^*=5-2=3$，查表 10-3 知 $x_2^*=2$，由 $s_3=s_2-x_2^*=3-2=1$，故 $x_3^*=s_3=1$。即得甲工厂分配 2 台，乙工厂分配 2 台，丙工厂分配1 台。

以上两个分配方案所得到的总赢利均为 21 万元。

在这个问题中，如果原设备的台数不是 5 台，而是 4 台或 3 台，用其他方法解时，往往要从头再算，但用动态规划解时，这些列出的表仍旧有用。只需要修改最后的表格，就可以得到：

当设备台数为 4 台时，最优分配方案为：$x_1^*=1,x_2^*=2,x_3^*=1$；或 $x_1^*=2,x_2^*=2,x_3^*=0$，总赢利为 17 万元。

当设备台数为 3 台时，最优分配方案为：$x_1^*=0,x_2^*=2,x_3^*=1$，总赢利为 14 万元。

这个例子是决策变量取离散值的一类分配问题。在实际中，如销售店分配问题，投资分配问题，货物分配问题等等，均属于这类分配问题。这种只将资源合理分配不考虑回收的问题，又称为**资源平行分配问题**。

在资源分配问题中，还有一种要考虑资源回收利用的问题，这里决策变量为连续值，故称为**资源连续分配问题**。这类分配问题一般叙述如下：

设有数量为 s_1 的某种资源，可投入 A 和 B 两种生产。第一年若以数量 u_1 投入生产 A，剩下的量 s_1-u_1 就投入生产 B，则可得收入为 $g(u_1)+h(s_1-u_1)$，其中 $g(u_1)$ 和 $h(u_1)$ 为已知函数，且 $g(0)=h(0)=0$。这种资源在投入 A、B 生产后，年终还可回收再投入生产。设年回收率分别为 $0<a<1$ 和 $0<b<1$，则在第一年生产后，回收的资源量合计为 $s_2=au_1+b(s_1-u_1)$。第二年再将资源数量 s_2 中的 u_2 和 s_2-u_2 分别再投入 A、B 两种生

产，则第二年又可得到收入为 $g(u_2)+h(s_2-u_2)$。如此继续进行 n 年，试问：应当如何决定每年投入 A 生产的资源量 $u_1, u_2, \cdots, u_n$，才能使总的收入最大？

此问题写成静态规划问题为

$$\max z=\{g(u_1)+h(s_1-u_1)+g(u_2)+h(s_2-u_2)+\cdots+g(u_n)+h(s_n-u_n)\}$$

$$\begin{cases} s_2 \quad =au_1+b(s_1-u_1) \\ s_3 \quad =au_2+b(s_2-u_2) \\ \qquad \vdots \\ s_{n+1}=au_n+b(s_n-u_n) \\ 0 \leqslant u_i \leqslant s_i, \quad i=1,2,\cdots,n \end{cases}$$

下面用动态规划方法来处理。

设 s_k 为状态变量，它表示在第 k 阶段（第 k 年）可投入 A、B 两种生产的资源量。

u_k 为决策变量，它表示在第 k 阶段（第 k 年）用于 A 生产的资源量，则 s_k-u_k 表示用于 B 生产的资源量。

状态转移方程为 $s_{k+1}=au_k+b(s_k-u_k)$

最优值函数 $f_k(s_k)$ 表示有资源量 s_k，从第 k 阶段至第 n 阶段采取最优分配方案进行生产后所得到的最大总收入。

因此可写出动态规划的逆推关系式为

$$\begin{cases} f_n(s_n)=\max\limits_{0\leqslant u_n\leqslant s_n}\{g(u_n)+h(s_n-u_n)\} \\ f_k(s_k)=\max\limits_{0\leqslant u_k\leqslant s_k}\{g(u_k)+h(s_k-u_k)+f_{k+1}[au_k+b(s_k-u_k)]\} \\ \qquad k=n-1,\cdots,2,1 \end{cases} \tag{10-1}$$

最后求出 $f_1(s_1)$ 即为所求问题的最大总收入。

例 2 机器负荷分配问题。某种机器可在高低两种不同的负荷下进行生产，设机器在高负荷下生产的产量函数为 $g=8u_1$，其中 u_1 为投入生产的机器数量，年完好率 $a=0.7$；在低负荷下生产的产量函数为 $h=5y$，其中 y 为投入生产的机器数量，年完好率为 $b=0.9$。

假定开始生产时完好的机器数量 $s_1=1\ 000$ 台，试问每年如何安排机器在高、低负荷下的生产，使在五年内生产的产品总产量最高。

先构造这个问题的动态规划模型：

设阶段序数 k 表示年度。

状态变量 s_k 为第 k 年度初拥有的完好机器数量，同时也是第 $k-1$ 年度末时的完好机器数量。

决策变量 u_k 为第 k 年度中分配高负荷下生产的机器数量，于是 s_k-u_k 为该年度中分配在低负荷下生产的机器数量。

这里 s_k 和 u_k 均取连续变量，它们的非整数值可以这样理解，如 $s_k=0.6$，就表示一台

机器在 k 年度中正常工作时间只占 6/10；$u_k=0.3$，就表示一台机器在该年度只有 3/10 的时间能在高负荷下工作。

状态转移方程为

$$s_{k+1}=au_k+b(s_k-u_k)=0.7u_k+0.9(s_k-u_k),\quad k=1,2,\cdots,5$$

k 段允许决策集合为 $D_k(s_k)=\{u_k|0\leqslant u_k\leqslant s_k\}$

设 $v_k(s_k,u_k)$ 为第 k 年度的产量，则

$$v_k=8u_k+5(s_k-u_k)$$

故指标函数为

$$V_{1,5}=\sum_{k=1}^{5}v_k(s_k,u_k)$$

令最优值函数 $f_k(s_k)$ 表示由资源量 s_k 出发，从第 k 年开始到第 5 年结束时所生产的产品的总产量最大值。因而有逆推关系式

$$\begin{cases}f_6(s_6)=0\\ f_k(s_k)=\max\limits_{u_k\in D_k(s_k)}\{8u_k+5(s_k-u_k)+f_{k+1}[0.7u_k+0.9(s_k-u_k)]\}\\ \qquad k=1,2,3,4,5\end{cases}$$

从第 5 年度开始，向前逆推计算。

当 $k=5$ 时，有

$$\begin{aligned}f_5(s_5)&=\max_{0\leqslant u_5\leqslant s_5}\{8u_5+5(s_5-u_5)+f_6[0.7u_5+0.9(s_5-u_5)]\}\\&=\max_{0\leqslant u_5\leqslant s_5}\{8u_5+5(s_5-u_5)\}\\&=\max_{0\leqslant u_5\leqslant s_5}\{3u_5+5s_5\}\end{aligned}$$

因 f_5 是 u_5 的线性单调增函数，故得最大解 $u_5^*=s_5$，相应的有 $f_5(s_5)=8s_5$。

当 $k=4$ 时，有

$$\begin{aligned}f_4(s_4)&=\max_{0\leqslant u_4\leqslant s_4}\{8u_4+5(s_4-u_4)+f_5[0.7u_4+0.9(s_4-u_4)]\}\\&=\max_{0\leqslant u_4\leqslant s_4}\{8u_4+5(s_4-u_4)+8[0.7u_4+0.9(s_4-u_4)]\}\\&=\max_{0\leqslant u_4\leqslant s_4}\{13.6u_4+12.2(s_4-u_4)\}\\&=\max_{0\leqslant u_4\leqslant s_4}\{1.4u_4+12.2s_4\}\end{aligned}$$

故得最大解，$u_4^*=s_4$，相应的有 $f_4(s_4)=13.6s_4$ 依此类推，可求得

$$u_3^*=s_3,\text{相应的 } f_3(s_3)=17.5s_3$$

$$u_2^*=0,\text{相应的 } f_2(s_2)=20.8s_2$$

$$u_1^*=0,\text{相应的 } f_1(s_1)=23.7s_1$$

因 $s_1=1\,000$，故 $f_1(s_1)=23\,700$(台)。

计算结果表明：最优策略为 $u_1^*=0,u_2^*=0,u_3^*=s_3,u_4^*=s_4,u_5^*=s_5$，即前两年应把年

初全部完好机器投入低负荷生产，后三年应把年初全部完好机器投入高负荷生产。这样所得的产量最高，其最高产量为 23 700 台。

在得到整个问题的最优指标函数值和最优策略后，还需反过来确定每年年初的状态，即从始端向终端递推计算出每年年初完好机器数。已知 $s_1=1\,000$ 台，于是可得

$$s_2=0.7u_1^*+0.9(s_1-u_1^*)=0.9s_1=900(\text{台})$$
$$s_3=0.7u_2^*+0.9(s_2-u_2^*)=0.9s_2=810(\text{台})$$
$$s_4=0.7u_3^*+0.9(s_3-u_3^*)=0.7s_3=567(\text{台})$$
$$s_5=0.7u_4^*+0.9(s_4-u_4^*)=0.7s_4=397(\text{台})$$
$$s_6=0.7u_5^*+0.9(s_5-u_5^*)=0.7s_5=278(\text{台})$$

上面所讨论的最优策略过程，始端状态 s_1 是固定的，终端状态 s_6 是自由的。由此所得出的最优策略称为始端固定终端自由的最优策略，实现的目标函数是五年里的产品总产量最高。

如果在终端也附加上一定的约束条件，如规定在第五年度结束时，完好的机器数量为 500 台（上面只有 278 台），问应如何安排生产，才能在满足这一终端要求的情况下产量最高？读者作为练习自己计算之。

下面进一步来讨论始端固定、终端自由的一般情况。

设有 n 个年度，在高、低负荷下生产的产量函数分别为 $g=cu_1$，$h=du_2$，其中 c、$d>0$，$c>d$。而年回收率分别为 a 和 b，$0<a<b<1$。试求出最优策略的一般关系式。显然，这时状态转移方程为

$$s_{k+1}=au_k+b(s_k-u_k),\qquad k=1,2,\cdots,n$$

k 段的指标函数为

$$v_k=cu_k+d(s_k-u_k),\qquad k=1,2,\cdots,n$$

令 $f_k(s_k)$ 表示由状态 s_k 出发，从第 k 年至第 n 年年末时所生产的产品的总产量最大值。

因而可写出逆推关系式为

$$\begin{cases} f_{n+1}(s_{n+1})=0 \\ f_k(s_k)=\max\limits_{0\leqslant u_k\leqslant s_k}\{cu_k+d(s_k-u_k)+f_{k+1}[au_k+b(s_k-u_k)]\} \\ \qquad k=1,2,\cdots,n \end{cases}\tag{10-2}$$

我们知道，在低负荷下生产的时间愈长，机器完好率愈高，但生产产量少。而在高负荷下生产产量会增加，但机器损坏大。这样，即使每台产量高，总起来看产量也不高。

从前面的数字计算可以看出，前几年一般是全部用于低负荷生产，后几年则全部用于高负荷生产，这样才产量最高。如果总共为 n 年，从低负荷转为高负荷生产的是第 t 年，$1\leqslant t\leqslant n$。即是说，从 1 至 $t-1$ 年在低负荷下生产，t 至 n 年在高负荷下生产。现在要分析 t 与系数 a、b、c、d 是什么关系。

从回收率看，$(b-a)$ 值愈大，表示在高负荷下生产时，机车损坏情况比在低负荷时严

重得多，因此 t 值应选大些。从产量看，$(c-d)$ 值愈大，表示在高负荷下生产较有利，故 t 应选小些。下面我们从式(10-2)这一基本方程出发来求出 t 与 $(b-a)$、$(c-d)$ 的关系。

令 $\zeta_k=u_k/s_k$。则在低负荷生产时有 $\zeta_k=0$，高负荷生产时有 $\zeta_k=1$。对第 n 段，我们有

$$\begin{aligned} f_n(s_n) &= \max_{0\leqslant u_n\leqslant s_n}\{cu_n+d(s_n-u_n)\} \\ &= \max_{0\leqslant u_n\leqslant s_n}\{(c-d)u_n+ds_n\} \\ &= \max_{0\leqslant \zeta_n\leqslant 1}\{(c-d)\zeta_n+d\}s_n \end{aligned}$$

由于 $c>d$，所以 ζ_n 应选 1 才能使 $f_n(s_n)$ 最大。也就是说，最后一年应全部投入高负荷生产。故 $f_n(s_n)=cs_n$。

对 $n-1$ 段，根据式(10-2)有

$$\begin{aligned} f_{n-1}(s_{n-1}) &= \max_{0\leqslant u_{n-1}\leqslant s_{n-1}}\{cu_{n-1}+d(s_{n-1}-u_{n-1})+f_n(s_n)\} \\ &= \max_{u_{n-1}}\{cu_{n-1}+d(s_{n-1}-u_{n-1})+cs_n\} \\ &= \max_{u_{n-1}}\{cu_{n-1}+d(s_{n-1}-u_{n-1})+c[au_{n-1}+b(s_{n-1}-u_{n-1})]\} \\ &= \max_{u_{n-1}}\{[(c-d)-c(b-a)]u_{n-1}+(d+cb)s_{n-1}\} \\ &= \max_{\zeta_{n-1}}\{[(c-d)-c(b-a)]\zeta_{n-1}+(d+cb)\}s_{n-1} \end{aligned} \tag{10-3}$$

因此，欲要满足上式极值关系的条件是当

$$c-d>c(b-a) \tag{10-4}$$

时，应取 $\zeta_{n-1}^*=1$。即 $n-1$ 年仍应全部在高负荷下生产。否则，当式(10-4)不满足时，应取 $\zeta_{n-1}^*=0$，即 $n-1$ 年应全部投入低负荷生产。

由前面知道，只要在第 k 年投入低负荷生产，那么递推计算结果必然是从第 1 年到第 k 年均为低负荷生产，即有 $\zeta_1^*=\zeta_2^*=\cdots=\zeta_k^*=0$。可见，算出 $\zeta_k^*=0$ 后，前几年就没有必要再计算了。故只需研究哪一年由低负荷转入高负荷生产，即 ζ 从那一年开始变为 1 就行。

根据这点，现只分析满足式(10-4)的情况。由于 $\zeta_{n-1}^*=1$，故式(10-3)变为

$$f_{n-1}(s_{n-1})=(c+ca)s_{n-1}=c(1+a)s_{n-1}$$

又由于 $s_{n-1}=au_{n-2}+b(s_{n-2}-u_{n-2})$，将它代入上式得

$$f_{n-1}(s_{n-1})=c(1+a)[(a-b)u_{n-2}+bs_{n-2}]$$

对 $n-2$ 段，由式(10-2)有

$$\begin{aligned} f_{n-2}(s_{n-2}) &= \max_{0\leqslant u_{n-2}\leqslant s_{n-2}}\{cu_{n-2}+d(s_{n-2}-u_{n-2})+f_{n-1}(s_{n-1})\} \\ &= \max_{u_{n-2}}\{cu_{n-2}+d(s_{n-2}-u_{n-2})+c(1+a)[(a-b)u_{n-2}+bs_{n-2}]\} \\ &= \max_{u_{n-2}}\{[(c-d)-c(1+a)(b-a)]u_{n-2}+[d+c(1+a)b]s_{n-2}\} \end{aligned}$$

$$=\max_{\zeta_{n-2}}\{[(c-d)-c(1+a)(b-a)]\zeta_{n-2}+d+cb(1+a)\}s_{n-2}$$

由此可知，只要满足极值条件式

$$c-d>c(1+a)(b-a)$$

就应选 $\zeta_{n-2}^*=1$，否则为 0，即应继续在高负荷下生产。

依此类推，如果转入高负荷下生产的是第 t 年，则由

$$f_t(s_t)=\max_{u_t}\{cu_t+d(s_t-u_t)+f_{t+1}(s_{t+1})\}$$

可以推出，应满足极值关系的条件必然是

$$\begin{cases}c-d>c(1+a+a^2+\cdots+a^{n-(t+1)})(b-a)\\c-d<c(1+a+a^2+\cdots+a^{n-t})(b-a)\end{cases}\qquad(10\text{-}5)$$

相应的有最优策略

$$\zeta_n^*=\zeta_{n-1}^*=\cdots=\zeta_t^*=1$$

$$\zeta_1^*=\zeta_2^*=\cdots=\zeta_{t-1}^*=0$$

它就是例 2 在始端固定终端自由情况下最优策略的一般结果。

从这个例子看到，应用动态规划，可以在不求出数值解的情况下，确定出最优策略的结构。

可见，只要知道了 a,b,c,d 四个值，就总可找到一个 t 值，满足式(10-5)，且

$$1\leqslant t\leqslant n-1$$

例如题中给定的 $a=0.7,b=0.9,c=8,d=5$，代入式(10-5)，应有

$$\frac{c-d}{c(b-a)}=\frac{1-d/c}{b-a}=\frac{1-5/8}{0.9-0.7}=\frac{3}{1.6}=1.875>(1+a)=1+0.7$$

可见，$n-t-1=5-t-1=1$，所以 $t=3$。即从第三年开始将全部机器投入高负荷生产，五年内总产量最高。

从这个例子可看到：当 $g(x)$、$h(x)$是线性函数时，问题是易于求解的，当它们为复杂函数时，求解就困难多了。但是，当 $g(x)$、$h(x)$为凸函数且 $h(0)=g(0)=0$ 时，可以证明在每个阶段上的最优决策总是取其上限或下限。因此，对于解的结构来说，它与$g(x)$、$h(x)$为线性的情况是类似的。

上面的讨论表明：当 x 在$[0,s_1]$上离散地变化时，利用递推关系式逐步计算或表格法而求出数值解。当 x 在$[0,s_1]$上连续地变化时，若 $g(x)$和 $h(x)$是线性函数或凸函数时，根据递推关系式运用解析法不难求出 $f_k(x)$和最优解；若 $g(x)$和 $h(x)$不是线性函数或凸函数时，一般来说，解析法不能奏效，那只好利用递推关系式(10-1)求其数值解。首先要把问题离散化，即把区间$[0,s_1]$进行分割，令 $x=0,\Delta,2\Delta,\cdots,m\Delta=s_1$。其 Δ 的大小，应根据计算精度和计算机容量等来确定。然后规定所有的 $f_k(x)$和决策变量只在这些分割点上取值。这样，递推关系式(10-1)便可写为

$$\begin{cases} f_n(i\Delta)=\max\limits_{0\leqslant j\leqslant i}\{g(j\Delta)+h(i\Delta-j\Delta)\} \\ f_k(i\Delta)=\max\limits_{0\leqslant j\leqslant i}\{g(j\Delta)+h(i\Delta-j\Delta)+f_{k+1}[a(j\Delta)+b(i\Delta-j\Delta)]\} \\ \qquad k=n-1,\cdots,2,1 \end{cases}$$

对 $i=0,1,\cdots,m$ 依次计算,可逐步求出 $f_n(i\Delta),f_{n-1}(i\Delta),\cdots,f_1(i\Delta)$ 及相应的最优决策,最后求得 $f_1(m\Delta)=f_1(s_1)$ 就是所求的最大总收入。这种离散化算法可以编成程序在计算机中计算。

10.1.2 二维资源分配问题

设有两种原料,数量各为 a 和 b 单位,需要分配用于生产 n 种产品。如果第一种原料以数量 x_i 为单位,第二种原料以数量 y_i 为单位,用于生产第 i 种产品,其收入为 $g_i(x_i,y_i)$。问应如何分配这两种原料于 n 种产品的生产使总收入最大?

此问题可写成静态规划问题

$$\begin{cases} \max\ [g_1(x_1,y_1)+g_2(x_2,y_2)+\cdots+g_n(x_n,y_n)] \\ x_1+x_2+\cdots+x_n=a \\ y_1+y_2+\cdots+y_n=b \\ x_i\geqslant 0,\quad y_i\geqslant 0,\quad (i=1,2,\cdots,n)\text{且为整数} \end{cases}$$

用动态规划方法来解,状态变量和决策变量要取二维的。

设状态变量 (x,y)

x——分配用于生产第 k 种产品至第 n 种产品的第一种原料的单位数量。

y——分配用于生产第 k 种产品至第 n 种产品的第二种原料的单位数量。

决策变量 (x_k,y_k)

x_k——分配给第 k 种产品用的第一种原料的单位数量。

y_k——分配给第 k 种产品用的第二种原料的单位数量。

状态转换关系:

$$\tilde{x}=x-x_k$$

$$\tilde{y}=y-y_k$$

式中 $\tilde{x}$ 和 $\tilde{y}$ 分别表示用来生产第 $k+1$ 种产品至第 n 种产品的第一种和第二种原料的单位数量。

允许决策集合

$$D_k(x,y)=\left\{u_k\ \middle|\ \begin{matrix}0\leqslant x_k\leqslant x\\ 0\leqslant y_k\leqslant y\end{matrix}\right\}$$

$f_k(x,y)$ 表示以第一种原料数量为 x 单位,第二种原料数量为 y 单位,分配用于生产第 k 种产品至第 n 种产品时所得到的最大收入。故可写出逆推关系为

$$\begin{cases} f_n(x,y)=g_n(x,y) \\ f_k(x,y)=\max\limits_{\substack{0\leqslant x_k\leqslant x \\ 0\leqslant y_k\leqslant y}}[g_k(x_k,y_k)+f_{k+1}(x-x_k,y-y_k)] \\ \qquad k=n-1,\cdots,1 \end{cases}$$

最后求得 $f_1(a,b)$ 即为所求问题的最大收入。

在实际问题中，由于 $g(x,y)$ 的复杂性，一般计算较难，常利用这个递推关系进行数值计算，并采用下面的方法进行降维和简化处理，以求得它的解或近似解。

1. 拉格朗日乘数法

引入拉格朗日乘数 λ，将二维分配问题化为

$$\max\ \{g_1(x_1,y_1)+g_2(x_2,y_2)+\cdots+g_n(x_n,y_n)-\lambda(y_1+y_2+\cdots+y_n)\}$$

满足条件

$$x_1+x_2+\cdots+x_n=a$$

$$x_i\geqslant 0,y_i\geqslant 0,\quad i=1,2,\cdots,n \text{ 且为整数}$$

其中 λ 作为一个固定的参数。

令

$$h_i(x_i)=h_i(x_i,\lambda)=\max_{y_i\geqslant 0}[g_i(x_i,y_i)-\lambda y_i]$$

$\left(\text{为了使此式有意义，可设 } \lim\limits_{y_i\to\infty}\dfrac{g_i(x_i,y_i)}{y_i}=0\right)$

于是问题变为

$$\max\ [h_1(x_1)+h_2(x_2)+\cdots+h_n(x_n)]$$

满足

$$x_1+x_2+\cdots+x_n=a,\quad x_i\geqslant 0 \text{ 且为整数}$$

这是一个一维分配问题，可用对一维的方法去求解。这里，由于 λ 是参数，因此，最优解 $\bar{x}_i$ 是参数 λ 的函数，相应的 $\bar{y}_i$ 也是 λ 的函数。即 $x_i=\bar{x}_i(\lambda)$，$y_i=\bar{y}_i(\lambda)$ 为其解。如果 $\sum\limits_{i=1}^{n}\bar{y}_i(\lambda)=b$，则可证明 $\{\bar{x}_i,\bar{y}_i\}$ 为原问题的最优解。如果 $\sum\limits_{i=1}^{n}\bar{y}_i(\lambda)\neq b$，我们将调整 λ 的值（利用插值法逐渐确定 λ），直到 $\sum\limits_{i=1}^{n}\overline{y_i}(\lambda)=b$ 满足为止。

这样的降维方法在理论上有保证在计算上是可行的，故对于高维的问题可以用上述拉格朗日乘数法的思想来降低维数。

2. 逐次逼近法

这是另一种降维方法，先保持一个变量不变，对另一个变量实现最优化，然后交替地固定，以迭代的形式反复进行，直到获得某种要求的程度为止。

先设 $x^{(0)}=\{x_1^{(0)},x_2^{(0)},\cdots,x_n^{(0)}\}$ 为满足 $\sum\limits_{i=1}^{n}x_i^{(0)}=a$ 的一个可行解，固定 x 在 $x^{(0)}$，先对 y 求解，则二维分配问题变为一维问题

$$\begin{cases}\max\left[g_1(x_1^{(0)},y_1)+g_2(x_2^{(0)},y_2)+\cdots+g_n(x_n^{(0)},y_n)\right]\\ y_1+y_2+\cdots+y_n=b,y_i\geqslant 0\text{ 且为整数}\end{cases}$$

可用对一维的方法来求解。设这解为 $y^{(0)}=\{y_1^{(0)},y_2^{(0)},\cdots,y_n^{(0)}\}$，然后再固定 y 为 $y^{(0)}$，对 x 求解，即

$$\begin{cases}\max\sum_{i=1}^{n}g_i(x_i,y_i^{(0)})\\ \sum_{i=1}^{n}x_i=a,\quad x_i\geqslant 0\text{ 且为整数}\end{cases}$$

设其解为 $x^{(1)}=\{x_1^{(1)},x_2^{(1)},\cdots,x_n^{(1)}\}$，再固定 x 为 $x^{(1)}$，对 y 求解，这样依次轮换下去得到一系列的解 $\{x^{(k)}\},\{y^{(k)}\}(k=0,1,\cdots)$

因为

$$\sum_{i=1}^{n}g_i(x_i^{(0)},y_i)\leqslant\sum_{i=1}^{n}g_i(x_i^{(0)},y_i^{(0)})\leqslant\sum_{i=1}^{n}g_i(x_i^{(1)},y_i^{(0)})$$

故函数值序列 $\left\{\sum_{i=1}^{n}g_i(x_i^{(k)},y_i^{(k)})\right\}$ 是单调上升的，但不一定收敛到绝对的最优解，一般只收敛到某一局部最优解。因此，在实际计算时，可选择几个初始点 $x^{(0)}$ 进行计算，然后从所得到的几个局部最优解中选出一个最好的。

3. 粗格子点法(疏密法)

在采用离散化的方法计算时，先将矩形定义域：$0\leqslant x\leqslant a,0\leqslant y\leqslant b$ 分成网格，然后在这些格子点上进行计算。如将 a、b 各分为 m_1 和 m_2 等份，则总共有 $(m_1+1)\cdot(m_2+1)$ 个格点，故对每个 k 值需要计算的 $f_k(x,y)$ 共有 $(m_1+1)\cdot(m_2+1)$ 个。因此，这里的计算量是相当大的。随着分点加多，格子点数也增多，那时的计算量将大得惊人。为了使计算可行，往往根据问题要求的精确度，采用粗格子点法逐步缩小区域来减少计算量。

粗格子点法是先用少数的格子点进行粗糙的计算，在求出相应的最优解后，再在最优解附近的小范围内进一步细分，并求在细分格子点上的最优解，如此继续细分下去直到满足要求为止。这方法也可能出现最优解“漏网”的情况，因此，应用此法时要结合对指标函数的特性进行分析。

逐次逼近法和粗格子点法虽有缺点，但在实际问题中，这两种方法的应用是比较广泛的。

10.1.3 固定资金分配问题

设有 n 个生产行业，都需要某两种资源。对于第 k 个生产行业，如果用第 1 种资源 x_k 和第 2 种资源 y_k 进行生产，可获得利润为 $r_k(x_k,y_k)$。若第 1 种资源的单位价格为 a，第 2 种资源的单位价格为 b，现有资金 Z。问应购买第 1 种资源多少单位(设为 X)，第 2 种资源多少单位(设为 Y)，分配到 n 个生产行业，使总利润最大?

此问题的数学模型可写为

$$\max\sum_{k=1}^{n} r_k(x_k, y_k)$$

$$\begin{cases}\sum_{k=1}^{n} x_k = X & x_k \text{ 为非负整数} \\ \sum_{k=1}^{n} y_k = Y & y_k \text{ 为非负整数} \\ aX + bY \leqslant Z \end{cases}$$

解决这个问题，可以从固定资金开始，找出所有满足 $aX+bY\leqslant Z$ 的 X 和 Y；然后将资源量最优地分配给 n 个生产行业，找出获利最大的。这当然是一个算法，但不是有效的。

如果我们把资源分配换算成资金分配，那样做要简单些。

(1) 把资源分配利润表换算成资金分配利润表，即将 $r_k(x_k, y_k)$ 换算成 $R_k(z)$，$z=0,1,\cdots,Z$。但必须注意，分配的资金应先使较贵的资源单位最大。

设有资金 $z(0\leqslant z\leqslant Z)$ 分配到第 k 个生产行业，则由 $Z=aX+bY$ 知，在给定 z 的情况下，若购买第 2 种资源 y_k 单位，则留下的资金只能购买第 1 种资源 x_k 单位，$x_k=\left[\dfrac{z-by_k}{a}\right]$，于是得到资金利润函数 $R_k(z)$ 为

$$R_k(z)=\max_{y_k=0,1,\cdots,(z/b)}\left\{r_k\left(\left[\frac{z-by_k}{a}\right], y_k\right)\right\}$$

式中 (z/b) 指以资金 z 购买第 2 种资源的最大单位数，$\left[\dfrac{z-by_k}{a}\right]$ 指以资金 z 购买了第 2 种资源 y_k 单位以后能购买第 1 种资源的最大单位数。

(2) 计算最优资金分配所获得最大利润。规定最优值函数 $f_k(z)$ 表示以总的资金 z 分配到 k 至 n 个生产行业可获得的最大利润。

则有逆推关系式

$$\begin{cases} f_k(z)=\max\limits_{z_k=0,1,\cdots,z}\left[R_k(z_k)+f_{k+1}(z-z_k)\right] \\ f_n(z)=R_n(z) \end{cases}$$

(3) 求出 $f_1(z)$，即为问题的解。这样，就把一个原含有两个状态变量的问题转化为只含有一个状态变量的问题。

10.2 生产与存储问题

在生产和经营管理中，经常遇到要合理地安排生产（或购买）与库存的问题，达到既要满足社会的需要，又要尽量降低成本费用。因此，正确制定生产（或采购）策略，确定不同

时期的生产量(或采购量)和库存量,以使总的生产成本费用和库存费用之和最小,这就是生产与存储问题的最优化目标。

10.2.1 生产计划问题

设某公司对某种产品要制定一项 n 个阶段的生产(或购买)计划。已知它的初始库存量为零,每阶段生产(或购买)该产品的数量有上限的限制;每阶段社会对该产品的需求量是已知的,公司保证供应;在 n 阶段末的终结库存量为零。问该公司如何制定每个阶段的生产(或采购)计划,从而使总成本最小。

设 d_k 为第 k 阶段对产品的需求量,x_k 为第 k 阶段该产品的生产量(或采购量),v_k 为第 k 阶段结束时的产品库存量。则有 $v_k=v_{k-1}+x_k-d_k$

$c_k(x_k)$表示第 k 阶段生产产品 x_k 时的成本费用,它包括生产准备成本 K 和产品成本 ax_k(其中 a 是单位产品成本)两项费用。即

$$c_k(x_k)=\begin{cases}0 & \text{当 } x_k=0\\ K+ax_k & \text{当 } x_k=1,2,\cdots,m\\ \infty & \text{当 } x_k>m\end{cases}$$

$h_k(v_k)$表示在第 k 阶段结束时有库存量 v_k 所需的存储费用。

故第 k 阶段的成本费用为 $c_k(x_k)+h_k(v_k)$。

m 表示每阶段最多能生产该产品的上限数。

因而,上述问题的数学模型为

$$\min g=\sum_{k=1}^{n}[c_k(x_k)+h_k(v_k)]$$

$$\begin{cases}v_0=0,v_n=0\\ v_k=\sum\limits_{j=1}^{k}(x_j-d_j)\geqslant 0 & k=2,\cdots,n-1\\ 0\leqslant x_k\leqslant m & k=1,2,\cdots,n\\ x_k \text{ 为整数} & k=1,2,\cdots,n\end{cases}$$

用动态规划方法来求解,把它看做一个 n 阶段决策问题。令

v_{k-1}为状态变量,它表示第 k 阶段开始时的库存量。

x_k 为决策变量,它表示第 k 阶段的生产量。

状态转移方程为

$$v_k=v_{k-1}+x_k-d_k \qquad k=1,2,\cdots,n$$

最优值函数 $f_k(v_k)$表示从第 1 阶段初始库存量为 0 到第 k 阶段末库存量为 v_k 时的最小总费用。

因此可写出顺序递推关系式为:

$$f_k(v_k)=\min_{0\leqslant x_k\leqslant \sigma_k}[c_k(x_k)+h_k(v_k)+f_{k-1}(v_{k-1})] \qquad k=1,\cdots,n$$

其中 $\sigma_k=\min(v_k+d_k,m)$。这是因为一方面每阶段生产的上限为 m；另一方面由于保证供应，故第 $k-1$ 阶段末的库存量 v_{k-1} 必须非负，即 $v_k+d_k-x_k\geqslant 0$，所以 $x_k\leqslant v_k+d_k$。

边界条件为 $f_0(v_0)=0$（或 $f_1(v_1)=\min\limits_{x_1=\sigma_1}[c_1(x_1)+h_1(v_1)]$），从边界条件出发，利用上面的递推关系式，对每个 k，计算出 $f_k(v_k)$ 中的 v_k 在 0 至 $\min\left[\sum\limits_{j=k+1}^{n}d_j, m-d_k\right]$ 之间的值，最后求得的 $f_n(0)$ 即为所求的最小总费用。

注 若每阶段生产产品的数量无上限的限制，则只要改变 $c_k(x_k)$ 和 σ_k 就行。即

$$c_k(x_k)=\begin{cases}0 & \text{当 } x_k=0\\ K+ax_k & \text{当 } x_k=1,2,\cdots\end{cases}$$

$$\sigma_k=v_k+d_k$$

对每个 k，需计算 $f_k(v_k)$ 中的 v_k 在 0 至 $\sum\limits_{j=k+1}^{n}d_j$ 之间的值。

例 3 某工厂要对一种产品制订今后四个时期的生产计划，据估计在今后四个时期内，市场对于该产品的需求量如表 10-5 所示。

表 10-5

时期(k)	1	2	3	4
需求量(d_k)	2	3	2	4

假定该厂生产每批产品的固定成本为 3 千元，若不生产就为 0；每单位产品成本为 1 千元；每个时期生产能力所允许的最大生产批量为不超过 6 个单位；每个时期末未售出的产品，每单位需付存储费 0.5 千元。还假定在第一个时期的初始库存量为 0，第四个时期之末的库存量也为 0。试问该厂应如何安排各个时期的生产与库存，才能在满足市场需要的条件下，使总成本最小。

解 用动态规划方法来求解，其符号含义与上面相同。

按四个时期将问题分为四个阶段。由题意知，在第 k 时期内的生产成本为

$$c_k(x_k)=\begin{cases}0 & \text{当 } x_k=0\\ 3+1\cdot x_k & \text{当 } x_k=1,2,\cdots,6\\ \infty & \text{当 } x_k>6\end{cases}$$

第 k 时期末库存量为 v_k 时的存储费用为

$$h_k(v_k)=0.5v_k$$

故第 k 时期内的总成本为 $c_k(x_k)+h_k(v_k)$

而动态规划的顺序递推关系式为

$$f_k(v_k)=\min_{0\leqslant x_k\leqslant\sigma_k}[c_k(x_k)+h_k(v_k)+f_{k-1}(v_k+d_k-x_k)],\quad k=1,2,3,4$$

式中 $\sigma_k=\min(v_k+d_k,6)$

和边界条件 $f_1(v_1)=\min\limits_{x_1=\min(v_1+d_1,6)}[c_1(x_1)+h_1(v_1)]$

当 $k=1$ 时，由

$$f_1(v_1)=\min_{x_1=\min(v_1+2,6)}[c_1(x_1)+h_1(v_1)]$$

对 v_1 在 0 至 $\min[\sum\limits_{j=2}^{4}d_j,m-d_1]=\min[9,6-2]=4$ 之间的值分别进行计算。

$v_1=0$ 时，　$f_1(0)=\min\limits_{x_1=2}[3+x_1+0.5\times0]=5$　　所以 $x_1=2$

$v_1=1$ 时，　$f_1(1)=\min\limits_{x_1=3}[3+x_1+0.5\times1]=6.5$　　所以 $x_1=3$

$v_1=2$ 时，　$f_1(2)=\min\limits_{x_1=4}[3+x_1+0.5\times2]=8$　　所以 $x_1=4$

同理得

$f_1(3)=9.5$　　所以 $x_1=5$

$f_1(4)=11$　　所以 $x_1=6$

当 $k=2$ 时，由

$$f_2(v_2)=\min_{0\leqslant x_2\leqslant\sigma_2}[c_2(x_2)+h_2(v_2)+f_1(v_2+3-x_2)]$$

其中 $\sigma_2=\min(v_2+3,6)$。对 v_2 在 0 至 $\min\left[\sum\limits_{j=3}^{4}d_j,6-3\right]=\min[6,3]=3$ 之间的值分别进行计算。从而有

$$f_2(0)=\min_{0\leqslant x_2\leqslant3}[c_2(x_2)+h_2(0)+f_1(3-x_2)]$$

$$=\min\begin{bmatrix}c_2(0)+h_2(0)+f_1(3)\\c_2(1)+h_2(0)+f_1(2)\\c_2(2)+h_2(0)+f_1(1)\\c_2(3)+h_2(0)+f_1(0)\end{bmatrix}=\min\begin{bmatrix}0+9.5\\4+8\\5+6.5\\6+5\end{bmatrix}=9.5\quad\text{所以 } x_2=0$$

$f_2(1)=\min\limits_{0\leqslant x_2\leqslant4}[c_2(x_2)+h_2(1)+f_1(4-x_2)]=11.5$　　所以 $x_2=0$

$f_2(2)=\min\limits_{0\leqslant x_2\leqslant5}[c_2(x_2)+h_2(2)+f_1(5-x_2)]=14$　　所以 $x_2=5$

$f_2(3)=\min\limits_{0\leqslant x_2\leqslant6}[c_2(x_2)+h_2(3)+f_1(6-x_2)]=15.5$　　所以 $x_2=6$

注　在计算 $f_2(2)$ 和 $f_2(3)$ 时，由于每个时期的最大生产批量为 6 单位，故 $f_1(5)$ 和 $f_1(6)$ 是没有意义的，就取 $f_1(5)=f_1(6)=\infty$，其余类推。

当 $k=3$ 时，由

$$f_3(v_3)=\min_{0\leqslant x_3\leqslant\sigma_3}[c_3(x_3)+h_3(v_3)+f_2(v_3+2-x_3)]$$

其中 $\sigma_3=\min(v_3+2,6)$。对 v_3 在 0 至 $\min[4,6-2]=4$ 之间的值分别进行计算，从而有

$f_3(0)=14$　　所以　$x_3=0$

$f_3(1)=16$　　所以　$x_3=0$ 或 3

$f_3(2)=17.5$　　所以　$x_3=4$

$$f_3(3)=19 \qquad 所以 \quad x_3=5$$

$$f_3(4)=20.5 \qquad 所以 \quad x_3=6$$

当 $k=4$ 时，因要求第 4 时期之末的库存量为 0，即 $v_4=0$，故有

$$f_4(0)=\min_{0\leqslant x_1\leqslant 4}[c_4(x_4)+h_4(0)+f_3(4-x_4)]$$

$$=\min\begin{bmatrix}c_4(0)+f_3(4)\\c_4(1)+f_3(3)\\c_4(2)+f_3(2)\\c_4(3)+f_3(1)\\c_4(4)+f_3(0)\end{bmatrix}=\min\begin{bmatrix}0+20.5\\4+19\\5+17.5\\6+16\\7+14\end{bmatrix}=20.5 \qquad 所以\ x_4=0$$

再按计算的顺序反推算，可找出每个时期的最优生产决策为：

$$x_1=5,x_2=0,x_3=6,x_4=0$$

其相应的最小总成本为 20.5 千元。

如果把上面例题中的有关数据列成表 10-6，然后分析这些数据，可找出一些规律性的东西。

表 10-6

阶段 i	0	1	2	3	4
需求量 d_i	—	2	3	2	4
生产量 x_i	—	5	0	6	0
库存量 v_i	0	3	0	4	0

由表中的数字可以看到，这样的库存问题有如下特征：

(1) 对每个 i，有 $v_{i-1}\cdot x_i=0$，$(i=1,2,3,4)$其中 $v_0=0$。

(2) 对于最优生产决策来说，它被裂解为两个子问题，一个是从第 1 阶段到第 2 阶段，另一个是从第 3 阶段到第 4 阶段。每个子问题的最优生产决策特别简单，它们的最小总成本之和就等于原问题的最小总成本。

这种现象不是偶然的，而是反映着这类库存问题数学模型的特征。研究这类问题，注意到(2)的规律，就可将计算量大量地减少。

如果对每个 i，都有 $v_{i-1}x_i=0$，则称该点的生产决策(或称一个策略 $x=x_1,\cdots,x_n$)具有**再生产点性质**(又称重生性质)。如果 $v_i=0$，则称阶段 i 为**再生产点**(又称重生点)。

由假设 $v_0=0$ 和 $v_n=0$，故阶段 0 和 n 是再生产点。可以证明：若库存问题的目标函数 $g(x)$在凸集合 S 上是凹函数(或凸函数)，则 $g(x)$在 S 的顶点上具有再生产点性质的最优策略。下面运用再生产点性质来求库存问题为凹函数的解。

设 $c(j,i)(j\leqslant i)$为阶段 j 到阶段 i 的总成本，给定 $j-1$ 和 i 是再生产点，并且阶段 j 到阶段 i 期间的产品全部由阶段 j 供给。则

$$c(j,i)=c_j\Big(\sum_{s=j}^{i}d_s\Big)+\sum_{s=j+1}^{i}c_s(0)+\sum_{s=j}^{i-1}h_s\Big(\sum_{t=s+1}^{i}d_t\Big) \tag{10-6}$$

根据两个再生点之间的最优策略,可以得到一个更有效的动态规划递推关系式。

设最优值函数 f_i 表示在阶段 i 末库存量 $v_i=0$ 时,从阶段 1 到阶段 i 的最小成本。则对应的递推关系式为

$$f_i=\min_{1\leqslant j\leqslant i}\left[f_{j-1}+c(j,i)\right] \quad i=1,2,\cdots,n \tag{10-7}$$

边界条件为

$$f_0=0 \tag{10-8}$$

为了确定最优生产决策,逐个计算 $f_1,f_2,\cdots,f_n$。则 $f_n(0)$ 为 n 个阶段的最小总成本。设 $j(n)$ 为计算 f_n 时,使式(10-7)右边最小的 j 值,即

$$f_n=\min_{1\leqslant j\leqslant n}\left[f_{j-1}+c(j,n)\right]=f_{j(n)-1}+c(j(n),n)$$

则从阶段 $j(n)$ 到阶段 n 的最优生产决策为

$$x_{j(n)}=\sum_{s=j(n)}^{n}d_s$$

$$x_s=0 \qquad 当\ s=j(n)+1,j(n)+2,\cdots,n\ 时$$

故阶段 $j(n)-1$ 为再生产点。为了进一步确定阶段 $j(n)-1$ 到阶段 1 的最优生产决策,记 $m=j(n)-1$,而 $j(m)$ 是在计算 f_m 时,使式(10-7)右边最小的 j 值,则从阶段 $j(m)$ 到阶段 $j(n)$ 的最优生产决策为

$$x_{j(m)}=\sum_{s=j(m)}^{m}d_s$$

$$x_s=0 \qquad 当\ s=j(m)+1,j(m)+2,\cdots,m\ 时$$

故阶段 $j(m)-1$ 为再生产点,其余依此类推。

例 4 利用再生产点性质解例 3。

解 因 $c_i(x_i)=\begin{cases}0 & x_i=0\\ 3+x_i & x_i=1,2,\cdots,6\\ \infty & x_i>6\end{cases}$。和 $h_i(v_i)=0.5v_i$

都是凹函数,故可利用再生产点性质来计算。

(1) 按(10-6)式计算 $c(j,i)$, $1\leqslant j\leqslant i$, $i=1,2,3,4$

$c(1,1)=c(2)+h(0)=5$

$c(1,2)=c(5)+h(3)=3+5+0.5\times3=9.5$

$c(1,3)=c(7)+h(5)+h(2)=\infty+0.5\times5+0.5\times2=\infty$

$c(1,4)=c(11)+h(9)+h(6)+h(4)=\infty$

$c(2,2)=c(3)+h(0)=6$ $\qquad$ $c(2,3)=c(5)+h(2)=9$

$c(2,4)=c(9)+h(6)+h(4)=\infty$ $\qquad$ $c(3,3)=c(2)+h(0)=5$

$c(3,4)=c(6)+h(4)=11$ $\qquad$ $c(4,4)=c(4)=7$

(2) 按式(10-7)和式(10-8)计算 f_i

$f_0=0$

$f_1=f_0+c(1,1)=0+5=5$ 所以 $j(1)=1$

$f_2=\min[f_0+c(1,2),f_1+c(2,2)]$

$=\min[0+9.5,5+6]=9.5$ 所以 $j(2)=1$

$f_3=\min[f_0+c(1,3),f_1+c(2,3),f_2+c(3,3)]$

$=\min[0+\infty,5+9,9.5+5]=14$ 所以 $j(3)=2$

$f_4=\min[f_0+c(1,4),f_1+c(2,4),f_2+c(3,4),f_3+c(4,4)]$

$=\min[0+\infty,5+\infty,9.5+11,14+7]$

$=20.5$ 所以 $j(4)=3$

(3) 找出最优生产决策

由 $j(4)=3$， 故 $x_3=d_3+d_4=6,x_4=0$

因 $m=j(4)-1=3-1=2$， 所以 $j(m)=j(2)=1$

故 $x_1=d_1+d_2=5,x_2=0$

所以最优生产决策为：$x_1=5,x_2=0,x_3=6,x_4=0$

相应的最小总成本为 20.5 千元。

还应指出的是，这种利用再生产点性质求解确定性需求不允许缺货的库存问题，可以推广到确定性需求在某些阶段上允许延迟交货的库存问题。

例 5 某车间需要按月在月底供应一定数量的某种部件给总装车间，由于生产条件的变化，该车间在各月份中生产每单位这种部件所需耗费的工时不同，各月份所生产的部件量于当月月底前全部要存入仓库以备后用。已知总装车间的各个月份的需求量以及在加工车间生产该部件每单位数量所需工时数如表 10-7 所示。

表 10-7

月份 k	0	1	2	3	4	5	6
需求量 d_k	0	8	5	3	2	7	4
单位工时 a_k	11	18	13	17	20	10	

设仓库容量限制为 $H=9$，开始库存量为 2，期终库存量为 0，需要制定一个半年的逐月生产计划，既使得满足需要和库容量的限制，又使得生产这种部件的总耗费工时数为最少。

解 按月份划分阶段，用 k 表示月份序号。

设状态变量 s_k 为第 k 段开始时(本段需求量送出之前，上段产品送入之后)部件库存量。

决策变量 u_k 为第 k 段内的部件生产量。

状态转移方程

$$s_{k+1}=s_k+u_k-d_k \qquad k=0,1,\cdots,6 \tag{10-9}$$

且

$$d_k \leqslant s_k \leqslant H \tag{10-10}$$

故允许决策集合为

$$D_k(s_k)=\{u_k: u_k \geqslant 0, d_{k+1} \leqslant s_k+u_k-d_k \leqslant H\} \tag{10-11}$$

最优值函数 $f_k(s_k)$ 表示在第 k 段开始的库存量为 s_k 时，从第 k 段至第 6 段所生产部件的最小累计工时数。

因而可写出逆推关系式为

$$\begin{cases} f_k(s_k)=\min\limits_{u_k \in D_k(s_k)} [a_k u_k+f_{k+1}(s_k+u_k-d_k)], \quad k=0,1,\cdots,6 \\ f_7(s_7)=0 \end{cases} \tag{10-12}$$

当 $k=6$ 时，因要求期终库存量为 0，即 $s_7=0$。因每月的生产是供应下月的需要，故第 6 个月不用生产，即 $u_6=0$。因此 $f_6(s_6)=0$，而由式(10-9)有 $s_6=d_6=4$

当 $k=5$ 时，由式(10-9)有 $s_6=s_5+u_5-d_5$，故 $u_5=11-s_5$

所以 $$f_5(s_5)=\min_{u_5=11-s_5}(a_5 u_5)=10(11-s_5)=110-10s_5$$

及最优解 $u_5^*=11-s_5$

当 $k=4$ 时，有

$$\begin{aligned} f_4(s_4) &= \min_{u_4 \in D_4(s_4)} [a_4 u_4+f_5(s_4+u_4-d_4)] \\ &= \min_{u_4} [20u_4+110-10(s_4+u_4-2)] \\ &= \min_{u_4} [10u_4-10s_4+130] \end{aligned}$$

其中 u_4 的允许决策集合 $D_4(s_4)$ 由式(10-11)确定为

由 $d_5 \leqslant s_4+u_4-d_4 \leqslant H$， 故有 $9-s_4 \leqslant u_4 \leqslant 11-s_4$

又 $u_4 \geqslant 0$，因而 $\max[0,9-s_4] \leqslant u_4 \leqslant 11-s_4$

而由式(10-10)知：$s_4 \leqslant 9$，所以 $D_4(s_4)$ 为

$$9-s_4 \leqslant u_4 \leqslant 11-s_4$$

故得 $f_4(s_4)=10(9-s_4)-10s_4+130=220-20s_4$ 及最优解 $u_4^*=9-s_4$

当 $k=3$ 时

$$\begin{aligned} f_3(s_3) &= \min_{u_3 \in D_3(s_3)} [a_3 u_3+f_4(s_3+u_3-d_3)] \\ &= \min_{u_3} [17u_3+220-20(s_3+u_3-3)] \\ &= \min_{u_3} [-3u_3-20s_3+280] \end{aligned}$$

由式(10-11)得 $D_3(s_3)$ 为 $\max[0,5-s_3] \leqslant u_3 \leqslant 12-s_3$

故得 $f_3(s_3)=244-17s_3$ 及最优解 $u_3^*=12-s_3$

当 $k=2$ 时

$$f_2(s_2)=\min_{u_2\in D_2(s_2)}\left[a_2u_2+f_3(s_2+u_2-d_2)\right]$$
$$=\min_{u_2}\left[-4u_2-17s_2+329\right]$$

其中 $D_2(s_2)$ 为 $\max[0,8-s_2]\leqslant u_2\leqslant 14-s_2$

故得 $f_2(s_2)=273-13s_2$ 及最优解 $u_2^*=14-s_2$

当 $k=1$ 时

$$f_1(s_1)=\min_{u_1\in D_1(s_1)}\left[a_1u_1+f_2(s_1+u_1-d_1)\right]$$
$$=\min_{u_1}\left[5u_1-13s_1+377\right]$$

其中 $D_1(s_1)$ 为 $13-s_1\leqslant u_1\leqslant 17-s_1$

故得 $f_1(s_1)=442-18s_1$ 及最优解 $u_1^*=13-s_1$

当 $k=0$ 时

$$f_0(s_0)=\min_{u_0\in D_0(s_0)}\left[a_0u_0+f_1(s_0+u_0-d_0)\right]$$
$$=\min_{u_0}\left[-7u_0+442-18s_0\right]$$

其中 $D_0(s_0)$ 为 $8-s_0\leqslant u_0\leqslant 9-s_0$

故得 $f_0(s_0)=379-11s_0$ 及最优解 $u_0^*=9-s_0$

因 $s_0=2$，所以 $f_0=357$ 和 $u_0^*=7$

再按计算顺序反推之，并结合式(10-9)的运算，即得各阶段的最优决策为：

$$u_0^*=7,\quad u_1^*=4,\quad u_2^*=9,\quad u_3^*=3,\quad u_4^*=0,\quad u_5^*=4$$

所以从 0～5 月的最优生产计划为：7,4,9,3,0,4，相应的最小总工时数为 357。

10.2.2 不确定性的采购问题

在实际问题中，还遇到某些多阶段决策过程，不是像前面所讨论的确定性那样，状态转移是完全确定的，而是出现了随机性因素，状态转移不能完全确定，它是按照某种已知的概率分布取值的。具有这种性质的多阶段决策过程就称为随机性的决策过程。同处理确定性问题类似，用动态规划的方法也可处理这种随机性问题，有的又称此为随机性动态规划。下面举一个简单的例子加以说明。

例 6 采购问题。某厂生产上需要在近五周内必须采购一批原料，而估计在未来五周内价格有波动，其浮动价格和概率已测得如表 10-8 所示。试求在哪一周以什么价格购入，使其采购价格的数学期望值最小，并求出期望值。

表 10-8

单价	概率
500	0.3
600	0.3
700	0.4

解 这里价格是一个随机变量,是按某种已知的概率分布取值的。用动态规划方法处理,按采购期限5周分为5个阶段,将每周的价格看做该阶段的状态。设

y_k——状态变量,表示第 k 周的实际价格。

x_k——决策变量,当 $x_k=1$,表示第 k 周决定采购;当 $x_k=0$,表示第 k 周决定等待。

y_{kE}——第 k 周决定等待,而在以后采取最优决策时采购价格的期望值。

$f_k(y_k)$——第 k 周实际价格为 y_k 时,从第 k 周至第5周采取最优决策所得的最小期望值。

因而可写出逆序递推关系式为

$$f_k(y_k)=\min\{y_k, y_{kE}\}, \qquad y_k\in s_k \tag{10-13}$$

$$f_5(y_k)=y_5, \qquad y_5\in s_5 \tag{10-14}$$

其中

$$s_k=\{500,600,700\}, \qquad k=1,2,3,4,5 \tag{10-15}$$

由 y_{kE} 和 $f_k(y_k)$ 的定义可知:

$$y_{kE}=Ef_{k+1}(y_{k+1})=0.3f_{k+1}(500)+0.3f_{k+1}(600)+0.4f_{k+1}(700) \tag{10-16}$$

并且得出最优决策为

$$x_k=\begin{cases}1(\text{采购}) & \text{当 } f_k(y_k)=y_k \\ 0(\text{等待}) & \text{当 } f_k(y_k)=y_{kE}\end{cases} \tag{10-17}$$

从最后一周开始,逐步向前递推计算,具体计算过程如下。

$k=5$ 时,因 $f_5(y_5)=y_5, y_5\in s_5$,故有

$$f_5(500)=500, \quad f_5(600)=600, \quad f_5(700)=700$$

即在第五周时,若所需的原料尚未买入,则无论市场价格如何,都必须采购,不能再等。

$k=4$ 时,由式(10-16)可知

$$\begin{aligned}y_{4E}&=0.3f_5(500)+0.3f_5(600)+0.4f_5(700)\\&=0.3\times500+0.3\times600+0.4\times700=610\end{aligned}$$

于是,由式(10-13)得

$$f_4(y_4)=\min_{y_4\in s_4}\{y_4, y_{4E}\}=\min_{y_4\in s_4}\{y_4, 610\}$$

$$=\begin{cases}500 & \text{若 } y_4=500\\ 600 & \text{若 } y_4=600\\ 610 & \text{若 } y_4=700\end{cases}$$

由式(10-17)可知,第四周的最优决策为

$$x_4=\begin{cases}1(\text{采购}) & \text{若 } y_4=500 \text{ 或 } 600\\ 0(\text{等待}) & \text{若 } y_4=700\end{cases}$$

同理求得

$$f_3(y_3)=\min_{y_3\in s_3}\{y_3,y_{3E}\}=\min_{y_3\in s_3}\{y_3,574\}$$

$$=\begin{cases}500 & 若\ y_3=500\\574 & 若\ y_3=600\ 或\ 700\end{cases}$$

所以

$$x_3=\begin{cases}1 & 若\ y_3=500\\0 & 若\ y_3=600\ 或\ 700\end{cases}$$

$$f_2(y_2)=\min_{y_2\in s_2}\{y_2,y_{2E}\}=\min_{y_2\in s_2}\{y_2,551.8\}$$

$$=\begin{cases}500 & 若\ y_2=500\\551.8 & 若\ y_2=600\ 或\ 700\end{cases}$$

所以

$$x_2=\begin{cases}1 & 若\ y_2=500\\0 & 若\ y_2=600\ 或\ 700\end{cases}$$

$$f_1(y_1)=\min_{y_1\in s_1}\{y_1,y_{1E}\}=\min_{y_1\in s_1}\{y_1,536.26\}$$

$$=\begin{cases}500 & 若\ y_1=500\\536.26 & 若\ y_1=600\ 或\ 700\end{cases}$$

所以

$$x_1=\begin{cases}1 & 若\ y_1=500\\0 & 若\ y_1=600\ 或\ 700\end{cases}$$

由上可知，最优采购策略为：在第一、二、三周时，若价格为 500 就采购，否则应该等待；在第四周时，价格为 500 或 600 应采购，否则就等待；在第五周时，无论什么价格都要采购。

依照上述最优策略进行采购时，价格(单价)的数学期望值为

$$\begin{aligned}&500\times0.3[1+0.7+0.7^2+0.7^3+0.7^3\times0.4]+\\&\quad 600\times0.3[0.7^3+0.4\times0.7^3]+700\times0.4^2\times0.7^3\\&=500\times0.80106+600\times0.14406+700\times0.05488\\&=525.382\approx525\end{aligned}$$

且

$$0.80106+0.14406+0.05488=1$$

10.3* 背包问题

有一个人带一个背包上山，其可携带物品重量的限度为 a 千克。设有 n 种物品可供他选择装入背包中，这 n 种物品编号为 $1,2,\cdots,n$。已知第 i 种物品每件重量为 w_i 千克，在上山过程中的作用(价值)是携带数量 x_i 的函数 $c_i(x_i)$。问此人应如何选择携带物品(各几件)，使所起作用(总价值)最大。这就是著名的背包问题，类似的问题有工厂里的下

料问题,运输中的货物装载问题,人造卫星内的物品装载问题等等。

设 x_i 为第 i 种物品的装入件数,则问题的数学模型为

$$\max f=\sum_{i=1}^{n} c_i(x_i)$$

$$\begin{cases}\sum_{i=1}^{n} w_i x_i \leqslant a \\ x_i \geqslant 0 \text{ 且为整数} \quad (i=1,2,\cdots,n)\end{cases}$$

它是一个整数规划问题。如果 x_i 只取 0 或 1,又称为 0—1 背包问题。下面用动态规划的方法来求解。

设按可装入物品的 n 种类划分为 n 个阶段。

状态变量 w 表示用于装第 1 种物品至第 k 种物品的总重量。

决策变量 x_k 表示装入第 k 种物品的件数。则状态转移方程为

$$\tilde{w}=w-x_k w_k$$

允许决策集合为

$$D_k(w)=\left\{x_k \mid 0 \leqslant x_k \leqslant \left[\frac{w}{w_k}\right]\right\}$$

最优值函数 $f_k(w)$ 是当总重量不超过 w 千克,背包中可以装入第 1 种到第 k 种物品的最大使用价值。

即 $$f_k(w)=\max_{\substack{\sum_{i=1}^{k} w_i x_i \leqslant w \\ x_i \geqslant 0,\text{且为整数}(i=1,2,\cdots,k)}} \sum_{i=1}^{k} c_i(x_i)$$

因而可写出动态规划的顺序递推关系为

$$f_1(w)=\max_{x_1=0,1,\cdots,[w/w_1]} c_1(x_1)$$

$$f_k(w)=\max_{x_k=0,1,\cdots,[w/w_k]} \{c_k(x_k)+f_{k-1}(w-w_k x_k)\} \quad 2 \leqslant k \leqslant n$$

然后,逐步计算出 $f_1(w),f_2(w),\cdots,f_n(w)$ 及相应的决策函数 $x_1(w),x_2(w),\cdots,x_n(w)$,最后得出的 $f_n(a)$ 就是所求的最大价值,其相应的最优策略由反推运算即可得出。

例 7 $$\max f=4x_1+5x_2+6x_3$$

$$\begin{cases}3x_1+4x_2+5x_3 \leqslant 10 \\ x_i \geqslant 0 \quad \text{且为整数}, \quad i=1,2,3\end{cases}$$

解 用动态规划方法来解,此问题变为求 $f_3(10)$。

而 $$f_3(10)=\max_{\substack{3x_1+4x_2+5x_3 \leqslant 10 \\ x_i \geqslant 0,\text{整数},i=1,2,3}} \{4x_1+5x_2+6x_3\}$$

$$=\max_{\substack{3x_1+4x_2 \leqslant 10-5x_3 \\ x_i \geqslant 0,\text{整数},i=1,2,3}} \{4x_1+5x_2+(6x_3)\}$$

$$=\max_{\substack{10-5x_3\geqslant 0\\ x_3\geqslant 0,\text{整数}}}\{6x_3+\max_{\substack{3x_1+4x_2\leqslant 10-5x_3\\ x_1\geqslant 0,x_2\geqslant 0,\text{整数}}}[4x_1+5x_2]\}$$

$$=\max_{x_3=0,1,2}\{6x_3+f_2(10-5x_3)\}$$

$$=\max\{0+f_2(10),6+f_2(5),12+f_2(0)\}$$

由此看到,要计算 $f_3(10)$,必须先计算出 $f_2(10),f_2(5),f_2(0)$。而

$$f_2(10)=\max_{\substack{3x_1+4x_2\leqslant 10\\ x_1\geqslant 0,x_2\geqslant 0,\text{整数}}}\{4x_1+5x_2\}$$

$$=\max_{\substack{3x_1\leqslant 10-4x_2\\ x_1\geqslant 0,x_2\geqslant 0,\text{整数}}}\{4x_1+(5x_2)\}$$

$$=\max_{\substack{10-4x_2\geqslant 0\\ x_2\geqslant 0,\text{整数}}}\{5x_2+\max_{\substack{3x_1\leqslant 10-4x_2\\ x_1\geqslant 0,\text{整数}}}(4x_1)\}$$

$$=\max_{x_2=0,1,2}\{5x_2+f_1(10-4x_2)\}$$

$$=\max\{f_1(10),5+f_1(6),10+f_1(2)\}$$

$$f_2(5)=\max_{\substack{3x_1+4x_2\leqslant 5\\ x_1\geqslant 0,x_2\geqslant 0,\text{整数}}}\{4x_1+5x_2\}=\max_{x_2=0,1}\{5x_2+f_1(5-4x_2)\}$$

$$=\max\{f_1(5),5+f_1(1)\}$$

$$f_2(0)=\max_{\substack{3x_1+4x_2\leqslant 0\\ x_1\geqslant 0,x_2\geqslant 0,\text{整数}}}\{4x_1+5x_2\}=\max_{x_2=0}\{5x_2+f_1(0-4x_2)\}=f_1(0)$$

为了要计算出 $f_2(10),f_2(5),f_2(0)$,必须先计算出 $f_1(10),f_1(6),f_1(5),f_1(2),f_1(1),f_1(0)$,一般地有

$$f_1(w)=\max_{\substack{3x_1\leqslant w\\ x_1\geqslant 0,\text{整数}}}(4x_1)=4\times(\text{不超过 } w/3 \text{ 的最大整数})=4\times[w/3]$$

相应的最优决策为 $x_1=[w/3]$,于是得到

$$f_1(10)=4\times 3=12 \qquad (x_1=3)$$
$$f_1(6)=4\times 2=8 \qquad (x_1=2)$$
$$f_1(5)=4\times 1=4 \qquad (x_1=1)$$
$$f_1(2)=4\times 0=0 \qquad (x_1=0)$$
$$f_1(1)=4\times 0=0 \qquad (x_1=0)$$
$$f_1(0)=4\times 0=0 \qquad (x_1=0)$$

从而

$$f_2(10)=\max\{f_1(10),5+f_1(6),10+f_1(0)\}$$
$$=\max\{12,5+8,10+0\}=13 \qquad (x_1=2,x_2=1)$$
$$f_2(5)=\max\{f_1(5),5+f_1(1)\}$$
$$=\max\{4,5+0\}=5 \qquad (x_1=0,x_2=1)$$
$$f_2(0)=f_1(0)=0 \qquad (x_1=0,x_2=0)$$

故最后得到

$$\begin{aligned} f_3(10) &= \max\{f_2(10), 6+f_2(5), 12+f_2(0)\} \\ &= \max\{13, 6+5, 12+0\} \\ &= 13 \qquad\qquad (x_1=2, x_2=1, x_3=0) \end{aligned}$$

所以，最优装入方案为 $x_1^*=2, x_2^*=1, x_3^*=0$，最大使用价值为 13。

注意

(1) 若使用计算机进行计算时，对 $f_1(w)$ 和 $f_2(w)(w=0,1,\cdots,10)$ 的值都应算出并存储起来以备用。

(2) 在实际问题中，当 a 不大时，为了计算的简便，可将单位重量 w_i 排成递减序列，然后逐个分析 x_i 能取值的可能性，并适当加以比较调整，再删掉某些可能性，这样能节省计算量。

(3) 当 n 很大时，就会产生存储量过大的困难。如果 $c_i(x_i)$ 都是线性函数 c_ix_i 的情形，可按单位重量的价值 $\rho_i=c_i/w_i(i=1,2,\cdots,n)$ 由小到大进行排列。设有 $\rho_1\leqslant\rho_2\leqslant\cdots\leqslant\rho_{n-1}\leqslant\rho_n$，则对于给定的可供装入重量 w，如果 $w<w_n$，背包内当然无法容纳第 n 种物品，即最优解中 $x_n^*=0$；如果 $w=kw_n$（k 为正整数），背包内必然仅含有第 n 种物品，即最优解为 $x_n^*=k, x_i^*=0(i\neq n)$；如果 $w>w_n$，且不是 w_n 的整数倍，这时背包容纳了第 n 种物品，甚至可能不是最优解。但可以找到一个粗略的估算公式：当 $w\geqslant\dfrac{\rho_n}{\rho_n-\rho_{n-1}}w_n$ 成立时，最优解中 x_n^* 一定大于或等于 1，即一定要装入第 n 种物品。

上面例子我们只考虑了背包重量的限制，它称为“一维背包问题”。如果还增加背包体积的限制为 b，并假设第 i 种物品每件的体积为 v_i 立方米，问应如何装使得总价值最大。这就是“二维背包问题”，它的数学模型为

$$\max f=\sum_{i=1}^{n} c_i(x_i)$$

$$\begin{cases} \sum\limits_{i=1}^{n} w_ix_i\leqslant a \\ \sum\limits_{i=1}^{n} v_ix_i\leqslant b \\ x_i\geqslant 0 \quad \text{且为整数}, \quad i=1,2,\cdots,n \end{cases}$$

用动态规划方法来解，其思想方法与一维背包问题完全类似，只是这时的状态变量是两个（重量和体积的限制），决策变量仍是一个（物品的件数）。设最优值函数 $f_k(w,v)$ 表示当总重量不超过 w 公斤，总体积不超过 v 立方米时，背包中装入第 1 种到第 k 种物品的最大使用价值。故

$$f_k(w,v)=\max_{\substack{\sum_{i=1}^{k} w_ix_i\leqslant w \\ \sum_{i=1}^{k} v_ix_i\leqslant v \\ x_i\geqslant 0,\text{整数}\, i=1,2,\cdots,k}} \sum_{i=1}^{k} c_i(x_i)$$

因而可写出顺序递推关系式为

$$f_k(w,v)=\max_{0\leqslant x_k\leqslant\min\left(\left[\frac{w}{w_k}\right],\left[\frac{v}{v_k}\right]\right)}\{c_k(x_k)+f_{k-1}(w-w_kx_k,v-v_kx_k)\}$$

$$1\leqslant k\leqslant n$$

$$f_0(w,v)=0$$

最后算出 $f_n(a,b)$即为所求的最大价值。

10.4* 复合系统工作可靠性问题

若某种机器的工作系统由 n 个部件串联组成，只要有一个部件失灵，整个系统就不能工作。为提高系统工作的可靠性，在每一个部件上均装有主要元件的备用件，并且设计了备用元件自动投入装置。显然，备用元件越多，整个系统正常工作的可靠性越大。但备用元件多了，整个系统的成本、重量、体积均相应加大，工作精度也降低。因此，最优化问题是在考虑上述限制条件下，应如何选择各部件的备用元件数，使整个系统的工作可靠性最大。

设部件 $i(i=1,2,\cdots,n)$上装有 u_i 个备用件时，它正常工作的概率为 $p_i(u_i)$。因此，整个系统正常工作的可靠性，可用它正常工作的概率衡量。即

$$P=\prod_{i=1}^{n}p_i(u_i)$$

设装一个部件 i 备用元件费用为 c_i，重量为 w_i，要求总费用不超过 c，总重量不超过 w，则这个问题有两个约束条件，它的静态规划模型为

$$\max P=\prod_{i=1}^{n}p_i(u_i)$$

$$\begin{cases}\sum_{i=1}^{n}c_iu_i\leqslant c\\ \sum_{i=1}^{n}w_iu_i\leqslant w\\ u_i\geqslant 0\text{ 且为整数}，\quad i=1,2,\cdots,n\end{cases}$$

这是一个非线性整数规划问题，因 u_i 要求为整数，且目标函数是非线性的。非线性整数规划是个较为复杂的问题，但是用动态规划方法来解还是比较容易的。

为了构造动态规划模型，根据有两个约束条件，就取二维状态变量，采用两个状态变量符号 x_k,y_k 来表达，其中

x_k——由第 k 个到第 n 个部件所容许使用的总费用。

y_k——由第 k 个到第 n 个部件所容许具有的总重量。

决策变量 u_k 为部件 k 上装的备用元件数，这里决策变量是一维的。

这样，状态转移方程为

$$x_{k+1}=x_k-u_kc_k$$
$$y_{k+1}=y_k-u_kw_k \qquad (1\leqslant k\leqslant n)$$

允许决策集合为

$$D_k(x_k,y_k)=\{u_k:\ 0\leqslant u_k\leqslant \min([x_k/c_k],[y_k/w_k])\}$$

最优值函数 $f_k(x_k,y_k)$ 为由状态 x_k 和 y_k 出发，从部件 k 到部件 n 的系统的最大可靠性。

因此，整机可靠性的动态规划基本方程为

$$\begin{cases} f_k(x_k,y_k)=\max\limits_{u_k\in D_k(x_k,y_k)}[p_k(u_k)f_{k+1}(x_k-c_ku_k,y_k-w_ku_k)] \\ \qquad\qquad k=n,n-1,\cdots,1 \\ f_{n+1}(x_{n+1},y_{n+1})=1 \end{cases}$$

边界条件为1，这是因为 x_{n+1}、y_{n+1} 均为零，装置根本不工作，故可靠性当然为1。最后计算得 $f_1(c,w)$ 即为所求问题的最大可靠性。

这个问题的特点是指标函数为连乘积形式，而不是连加形式，但仍满足可分离性和递推关系；边界条件为1而不是零。它们是由研究对象的特性所决定的。另外，这里可靠性 $p_i(u_i)$ 是 u_i 的严格单调上升函数，而且 $p_i(u_i)\leqslant 1$。

在这个问题中，如果静态模型的约束条件增加为三个，例如要求总体积不许超过 v，则状态变量就要取为三维的 (x_k,y_k,z_k)。它说明静态规划问题的约束条件增加时，对应的动态规划的状态变量维数也需要增加，而决策变量维数可以不变。

例8 某厂设计一种电子设备，由三种元件 D_1、D_2、D_3 组成。已知这三种元件的价格和可靠性如表10-9所示，要求在设计中所使用元件的费用不超过105元。试问应如何设计使设备的可靠性达到最大(不考虑重量的限制)。

表 10-9

元　件	单位/元	可靠性
D_1	30	0.9
D_2	15	0.8
D_3	20	0.5

解 按元件种类划分为三个阶段，设状态变量 s_k 表示能容许用在 D_k 元件至 D_3 元件的总费用；决策变量 x_k 表示在 D_k 元件上的并联个数；p_k 表示一个 D_k 元件正常工作的概率，则 $(1-p_k)^{x_k}$ 为 x_k 个 D_k 元件不正常工作的概率。令最优值函数 $f_k(s_k)$ 表示由状态 s_k 开始从 D_k 元件至 D_3 元件组成的系统的最大可靠性。因而有

$$f_3(s_3)=\max_{1\leqslant x_3\leqslant [s_3/20]}[1-(0.5)^{x_3}]$$

$$f_2(s_2)=\max_{1\leqslant x_2\leqslant [s_2/15]}\{[1-(0.2)^{x_2}]f_3(s_2-15x_2)\}$$

$$f_1(s_1)=\max_{1\leqslant x_1\leqslant[s_1/30]}\{[1-(0.1)^{x_1}]f_2(s_1-30x_1)\}$$

由于 $s_1=105$，故此问题为求出 $f_1(105)$ 即可。

而 $f_1(105)=\max\limits_{1\leqslant x_1\leqslant 3}\{[1-(0.1)^{x_1}]f_2(105-30x_1)\}$

$=\max\{0.9f_2(75),0.99f_2(45),0.999f_2(15)\}$

但 $f_2(75)=\max\limits_{1\leqslant x_2\leqslant 4}\{[1-(0.2)^{x_2}]f_3(75-15x_2)\}$

$=\max\{0.8f_3(60),0.96f_3(45),0.992f_3(30),0.9984f_3(15)\}$

可是 $f_3(60)=\max\limits_{1\leqslant x_3\leqslant 3}[1-(0.5)^{x_3}]=\max\{0.5,0.75,0.875\}=0.875$

$f_3(45)=\max\{0.5,0.75\}=0.75$

$f_3(30)=0.5$

$f_3(15)=0$

所以 $f_2(75)=\max\{0.8\times0.875,0.96\times0.75,0.992\times0.5,0.9984\times0\}$

$=\max\{0.7,0.72,0.496,0\}=0.72$

同理 $f_2(45)=\max\{0.8f_3(30),0.96f_3(15)\}$

$=\max\{0.4,0\}=0.4$

$f_2(15)=0$

故 $f_1(105)=\max\{0.9\times0.72,0.99\times0.4,0.999\times0\}$

$=\max\{0.648,0.396,0\}=0.648$

从而求得 $x_1=1,x_2=2,x_3=2$ 为最优方案，即 D_1 元件用 1 个，D_2 元件用 2 个，D_3 元件用 2 个。其总费用为 100 元，可靠性为 0.648。

10.5 排序问题

设有 n 个工件需要在机床 A、B 上加工，每个工件都必须经过先 A 而后 B 的两道加工工序（见图 10-1）。以 a_i、b_i 分别表示工件 $i(1\leqslant i\leqslant n)$ 在 A、B 上的加工时间。问应如何在两机床上安排各工件加工的顺序，使在机床 A 上加工第一个工件开始到在机床 B 上将最后一个工件加工完为止，所用的加工总时间最少？

加工工件在机床 A 上有加工顺序问题，在机床 B 上也有加工顺序问题。它们在 A、B 两台机床上加工工件的顺序是可以不同的。当机床 B 上的加工顺序与机床 A 不同时，意味着在机床 A 上加工完毕的某些工件，不能在机床 B 上立即加工，而是要等到另一个或一些工件加工完毕之后才能加工。这样，使机床 B 的等待加工时间加长，从而使总的加工时间加长了。可以证

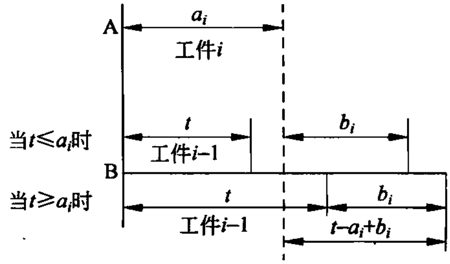


图 10-1

明：最优加工顺序在两台机床上可同时产生。因此，最优排序方案只能在机床 A、B 上加工顺序相同的排序中去寻找。即使如此，所有可能的方案仍有 $n!$ 个，这是一个不小的数，用穷举法是不现实的。下面用动态规划方法来研究同顺序两台机床加工 n 个工件的排序问题。

当加工顺序取定之后，工件在 A 上加工时没有等待时间，而在 B 上则常常等待。因此，寻求最优排序方案只有尽量减少在 B 上等待加工的时间，才能使总加工时间最短。设第 i 个工件在机床 A 上加工完毕以后，在 B 上要经过若干时间才能加工完，故对同一个工件来说，在 A、B 上总是出现加工完毕的时间差，我们以它来描述加工状态。

现在，我们以在机床 A 上更换工件的时刻作为时段。以 X 表示在机床 A 上等待加工的按取定顺序排列的工件集合。以 x 表示不属于 X 的在 A 上最后加工完的工件。以 t 表示在 A 上加工完 x 的时刻算起到 B 上加工完 x 所需的时间。这样，在 A 上加工完一个工件之后，就有 (X,t) 与之对应。

选取 (X,t) 作为描述机床 A、B 在加工过程中的状态变量。这样选取状态变量，则当 X 包含有 s 个工件时，过程尚有 s 段，其时段数已隐含在状态变量之中，因而，指标最优值函数只依赖于状态而不明显依赖于时段数。

令 $f(X,t)$ 为由状态 (X,t) 出发，对未加工的工件采取最优加工顺序后，将 X 中所有工件加工完所需时间。

$f(X,t,i)$ 为由状态 (X,t) 出发，在 A 上加工工件 i，然后再对以后的加工工件采取最优顺序后，把 X 中工件全部加工完所需要的时间。

$f(X,t,i,j)$ 为由状态 (X,t) 出发，在 A 上相继加工工件 i 与 j 后，对以后加工的工件采取最优顺序后，将 X 中的工件全部加工完所需要的时间。

因而，不难得到

$$f(X,t,i)=\begin{cases} a_i+f(X/i,t-a_i+b_i) & \text{当 } t\geqslant a_i \text{ 时} \\ a_i+f(X/i,b_i) & \text{当 } t\leqslant a_i \text{ 时} \end{cases}$$

式中状态 t 的转换关系参看示意图 10-1。

记
$$z_i(t)=\max\ (t-a_i,0)+b_i$$

上式就可合并写成

$$f(X,t,i)=a_i+f[X/i,z_i(t)]$$

其中 X/i 表示在集合 X 中去掉工件 i 后剩下的工件集合。

由定义，可得

$$f(X,t,i,j)=a_i+a_j+f\ [X/\{i,j\},z_{ij}(t)]$$

其中 $z_{ij}(t)$ 是在机床 A 上从 X 出发相继加工工件 i、j，并从它将 j 加工完的时刻算起，至在 B 上相继加工工件 i、j 并将工件加工完所需时间。故 $(X/\{i,j\},z_{ij}(t))$ 是在 A 加工 i、j 后所形成的新状态。即在机床 A 上加工 i、j 后由状态 (X,t) 转移到状态 $(X/\{i,j\},z_{ij}(t))$。

仿照 $z_i(t)$ 的定义，以 $X/\{i,j\}$ 代替 $X/\{i\}$，$z_i(t)$ 代替 t，a_j 代替 a_i，b_j 代替 b_i，则可得

$$z_{ij}(t)=\max\left[z_i(t)-a_j,0\right]+b_j$$

故

$$\begin{aligned}z_{ij}(t)&=\max\left[\max(t-a_i,0)+b_i-a_j,0\right]+b_j\\&=\max\left[\max(t-a_i-a_j+b_i,b_i-a_j),0\right]+b_j\\&=\max\left[t-a_i-a_j+b_i+b_j,b_i+b_j-a_j,b_j\right]\end{aligned}$$

将 i、j 对调，可得

$$f(X,t,j,i)=a_i+a_j+f\left[X/\{i,j\},z_{ji}(t)\right]$$

$$z_{ji}(t)=\max\left[t-a_i-a_j+b_i+b_j,b_i+b_j-a_i,b_i\right]$$

由于 $f(X,t)$ 为 t 的单调上升函数，故当 $z_{ij}(t)\leqslant z_{ji}(t)$ 时，有

$$f(X,t,i,j)\leqslant f(X,t,j,i)$$

因此，不管 t 为何值，当 $z_{ij}(t)\leqslant z_{ji}(t)$ 时，工件 i 放在工件 j 之前加工可以使总的加工时间短些。而由 $z_{ij}(t)$ 和 $z_{ji}(t)$ 的表示式可知，这只需要下面不等式成立就行。即

$$\max(b_i+b_j-a_j,b_j)\leqslant\max(b_i+b_j-a_i,b_i)$$

将上不等式两边同减去 b_i 与 b_j，得

$$\max(-a_j,-b_i)\leqslant\max(-a_i,-b_j)$$

即有

$$\min(a_i,b_j)\leqslant\min(a_j,b_i)$$

这个条件就是工件 i 应该排在工件 j 之前的条件。即对于从头到尾的最优排序而言。它的所有前后相邻接的两个工件所组成的对，都必须满足上不等式。根据这个条件，得到最优排序的规则如下：

(1) 先作工件的加工时间的工时矩阵

$$\boldsymbol{M}=\begin{pmatrix}a_1 & a_2 & \cdots & a_n\\ b_1 & b_2 & \cdots & b_n\end{pmatrix}$$

(2) 在工时矩阵 $\boldsymbol{M}$ 中找出最小元素（若最小的不止一个，可任选其一）；若它在上行，则将相应的工件排在最前位置；若它在下行，则将相应的工件排在最后位置。

(3) 将排定位置的工件所对应的列从 $\boldsymbol{M}$ 中划掉，然后对余下的工件重复按(2)进行。但那时的最前位置（或最后位置）是在已排定位置的工件之后（或之前）。如此继续下去，直至把所有工件都排完为止。

这个同顺序两台机床加工 n 个工件的最优排序规则，是 Johnson 在 1954 年提出的。概括起来说，它的基本思路是：尽量减少在机床 B 上等待加工的时间。因此，把在机床 B 上加工时间长的工件先加工，在 B 上加工时间短的工件后加工。

例 9　设有 5 个工件需在机床 A、B 上加工，加工的顺序是先 A 后 B，每个工件所需加工时间（单位：小时）如表 10-10 所示。问如何安排加工顺序，使机床连续加工完所有工件的加工总时间最少？并求出总加工时间。

表 10-10

工件号码	加工时间/h	
	机床 A	机床 B
1	3	6
2	7	2
3	4	7
4	5	3
5	7	4

解 工件的加工工时矩阵为

$$\boldsymbol{M}=\begin{bmatrix}3 & 7 & 4 & 5 & 7\\6 & 2 & 7 & 3 & 4\end{bmatrix}$$

根据最优排序规则，故最优加工顺序为

$$1\to3\to5\to4\to2$$

总加工时间为 28 小时。

10.6 设备更新问题

在工业和交通运输企业中，经常碰到设备陈旧或部分损坏需要更新的问题。从经济上来分析，一种设备应该用多少年后进行更新为最恰当，即更新的最佳策略应该如何，从而使在某一时间内的总收入达到最大(或总费用达到最小)。

现以一台机器为例，随着使用年限的增加，机器的使用效率降低，收入减少，维修费用增加。而且机器使用年限越长，它本身的价值就越小，因而更新时所需的净支出费用就愈多。设：

$I_j(t)$——在第 j 年机器役龄为 t 年的一台机器运行所得的收入。

$O_j(t)$——在第 j 年机器役龄为 t 年的一台机器运行时所需的运行费用。

$C_j(t)$——在第 j 年机器役龄为 t 年的一台机器更新时所需更新净费用。

α——折扣因子($0\leqslant\alpha\leqslant1$)，表示一年以后的单位收入的价值视为现年的 α 单位。

T——在第一年开始时，正在使用的机器的役龄。

n——计划的年限总数。

$g_j(t)$——在第 j 年开始使用一个役龄为 t 年的机器时，从第 j 年至第 n 年内的最佳收入。

$x_j(t)$——给出 $g_j(t)$ 时，在第 j 年开始时的决策(保留或更新)。

为了写出递推关系式，先从两方面分析问题。若在第 j 年开始时购买了新机器，则从第 j 年至第 n 年得到的总收入应等于在第 j 年中由新机器获得的收入，减去在第 j 年中

的运行费用,减去在第 j 年开始时役龄为 t 年的机器的更新净费用,加上在第 $j+1$ 年开始使用役龄为 1 年的机器从第 $j+1$ 年至第 n 年的最佳收入;若在第 j 年开始时继续使用役龄为 t 年的机器,则从第 j 年至第 n 年的总收入应等于在第 j 年由役龄为 t 年的机器得到的收入,减去在第 j 年中役龄为 t 年的机器的运行费用,加上在第 $j+1$ 年开始使用役龄为 $t+1$ 年的机器从第 $j+1$ 年至第 n 年的最佳收入。然后,比较它们的大小,选取大的,并相应得出是更新还是保留的决策。

将上面这段话写成数学形式,即得递推关系式为

$$g_j(t)=\max\begin{bmatrix}\text{R}:\ I_j(0)-O_j(0)-C_j(t)+\alpha g_{j+1}(1)\\ \text{K}:\ I_j(t)-O_j(t)+\alpha g_{j+1}(t+1)\end{bmatrix}$$

$$(j=1,2,\cdots,n\qquad t=1,2,\cdots,j-1,j+T-1)$$

其中“K”是 Keep 的缩写,表示保留使用;“R”是 Replacement 的缩写,表示更新机器。

由于研究的是今后 n 年的计划,故还要求

$$g_{n+1}(t)=0$$

对于 $g_1(\cdot)$来说,允许的 t 值只能是 T。因为当进入计划过程时,机器必然已使用了 T 年。

应指出的是:这里研究的设备更新问题,是以机龄作为状态变量,决策是保留和更新两种。但它可推广到多维情形,如还考虑对使用的机器进行大修作为一种决策,那时所需的费用和收入,不仅取决于机龄和购置的年限,也取决于上次大修后的时间。因此,必须使用两个状态变量来描述系统的状态,其过程与此类似。

例 10 假设 $n=5,\alpha=1,T=1$,其有关数据如表 10-11 所示。试制定 5 年中的设备更新策略,使在 5 年内的总收入达到最大。

表 10-11

项目	第一年					第二年				第三年			第四年		第五年	期前				
	机龄																			
	0	1	2	3	4	0	1	2	3	0	1	2	0	1	0	1	2	3	4	5
收入	22	21	20	18	16	27	25	24	22	29	26	24	30	28	32	18	16	16	14	14
运行费用	6	6	8	8	10	5	6	8	9	5	5	6	4	5	4	8	8	9	9	10
更新费用	27	29	32	34	37	29	31	34	36	31	32	33	32	33	34	32	34	36	36	38

解 先解释符号的意思。因第 j 年开始机龄为 t 年的机器,其制造年序应为 $j-t$ 年,因此,$I_5(0)$为第五年新产品的收入,故 $I_5(0)=32$。$I_3(2)$为第一年的产品其机龄为 2 年的收入,故 $I_3(2)=20$。同理 $O_5(0)=4,O_3(2)=8$。而 $C_5(1)$是第 5 年机龄为 1 年的机器(应为第四年的产品)的更新费用,故 $C_5(1)=33$。同理 $C_5(2)=33,C_3(1)=31$,其

余类推。

当 $j=5$ 时，由于设 $T=1$，故从第5年开始计算时，机器使用了1、2、3、4、5年，则递推关系式为

$$g_5(t)=\max\begin{bmatrix}R: I_5(0)-O_5(0)-C_5(t)+1\cdot g_6(1)\\K: I_5(t)-O_5(t)+1\cdot g_6(t+1)\end{bmatrix}$$

因此

$$g_5(1)=\max\begin{bmatrix}R: 32-4-33+0=-5\\K: 28-5+0=23\end{bmatrix}=23 \quad 所以\ x_5(1)=K$$

$$g_5(2)=\max\begin{bmatrix}R: 32-4-33+0=-5\\K: 24-6+0=18\end{bmatrix}=18 \quad 所以\ x_5(2)=K$$

同理 $g_5(3)=13, x_5(3)=K$; $g_5(4)=6, x_5(4)=K$

$g_5(5)=4, x_5(5)=K$

当 $j=4$ 时，递推关系为

$$g_4(t)=\max\begin{bmatrix}R: I_4(0)-O_4(0)-C_4(t)+g_5(1)\\K: I_4(t)-O_4(t)+g_5(t+1)\end{bmatrix}$$

故

$$g_4(1)=\max\begin{bmatrix}R: 30-4-32+23=17\\K: 26-5+18=39\end{bmatrix}=39 \quad 所以\ x_4(1)=K$$

同理 $g_4(2)=29, x_4(2)=K$; $g_4(3)=16, x_4(3)=K$

$g_4(4)=13, x_4(4)=R$

当 $j=3$ 时，有

$$g_3(t)=\max\begin{bmatrix}R: I_3(0)-O_3(0)-C_3(t)+g_4(1)\\K: I_3(t)-O_3(t)+g_4(t+1)\end{bmatrix}$$

故

$$g_3(1)=\max\begin{bmatrix}R: 29-5-31+39=32\\K: 25-6+29=48\end{bmatrix}=48 \quad 所以\ x_3(1)=K$$

同理 $g_3(2)=31, x_3(2)=R$; $g_3(3)=27, x_3(3)=R$

当 $j=2$ 时，有

$$g_2(t)=\max\begin{bmatrix}R: I_2(0)-O_2(0)-C_2(t)+g_3(1)\\K: I_2(t)-O_2(t)+g_3(t+1)\end{bmatrix}$$

故

$$g_2(1)=\max\begin{bmatrix}R: 27-5-29+48=41\\K: 21-6+31=46\end{bmatrix}=46 \quad 所以\ x_2(1)=K$$

$$g_2(2)=\max\begin{bmatrix}R: 27-5-34+48=36\\K: 16-8+27=35\end{bmatrix}=36 \quad x_2(2)=R$$

当 $j=1$ 时，有

$$g_1(t)=\max\begin{bmatrix}R: I_1(0)-O_1(0)-C_1(t)+g_2(1)\\K: I_1(t)-O_1(t)+g_2(t+1)\end{bmatrix}$$

故 $$g_1(1)=\max\begin{bmatrix}R: 22-6-32+46=30\\K: 18-8+36=46\end{bmatrix}=46 \quad 所以\ x_1(1)=K$$

最后，根据上面计算过程反推之，可求得最优策略如表 10-12 所示，相应的最佳收益为 46 单位。

表 10-12

年	机　龄	最佳策略
1	1	K
2	2	R
3	1	K
4	2	K
5	3	K

10.7* 货郎担问题

货郎担问题在运筹学里是一个著名的命题，有一个串村走户卖货郎，他从某个村庄出发，通过若干个村庄一次且仅一次，最后仍回到原出发的村庄，问应如何选择行走路线，能使总的行程最短。类似的问题有旅行路线问题，应如何选择行走路线，使总路程最短或费用最少。

现在把问题一般化。设有 n 个城市，以 $1,2,\cdots,n$ 表示之。d_{ij} 表示从 i 城到 j 城的距离。一个推销员从城市 1 出发到其他每个城市去一次且仅仅是一次，然后回到城市 1。问他如何选择行走的路线，使总的路程最短。这个问题属于组合最优化问题，当 n 不太大时，利用动态规划方法求解是很方便的。

由于规定推销员是从城市 1 开始的，设推销员走到 i 城，记

$N_i=\{2,3,\cdots,i-1,i+1,\cdots,n\}$ 表示由 1 城到 i 城的中间城市集合。

S 表示到达 i 城之前中途所经过的城市的集合，则有　$S\subseteq N_i$

因此，可选取 (i,S) 作为描述过程的状态变量，决策为由一个城市走到另一个城市，并定义最优值函数 $f_k(i,S)$ 为从 1 城开始经由 k 个中间城市的 S 集合到 i 城的最短路线的距离，则可写出动态规划的递推关系为

$$f_k(i,S)=\min_{j\in s}[f_{k-1}(j,S\backslash\{j\})+d_{ji}]$$

$$(k=1,2,\cdots,n-1。i=2,3,\cdots,n。S\subseteq N_i)$$

边界条件为　$f_0(i,\Phi)=d_{1i}$

$P_k(i,S)$ 为最优决策函数，它表示从 1 城开始经 k 个中间城市的 S 集合到 i 城的最短路线上紧挨着 i 城前面的那个城市。

例 11 求解四个城市旅行推销员问题,其距离矩阵如表 10-13 所示。当推销员从 1 城出发,经过每个城市一次且仅一次,最后回到 1 城,问按怎样的路线走,使总的行程距离最短。

表 10-13

j	i			
	1	2	3	4
	距离			
1	0	8	5	6
2	6	0	8	5
3	7	9	0	5
4	9	7	8	0

解 由边界条件可知

$$f_0(2,\Phi)=d_{12}=8,\quad f_0(3,\Phi)=d_{13}=5,\quad f_0(4,\Phi)=d_{14}=6$$

当 $k=1$ 时,即从 1 城开始,中间经过一个城市到达 i 城的最短距离是

$$f_1(2,\{3\})=f_0(3,\Phi)+d_{32}=5+9=14$$

$$f_1(2,\{4\})=f_0(4,\Phi)+d_{42}=6+7=13$$

$$f_1(3,\{2\})=8+8=16 \qquad f_1(3,\{4\})=6+8=14$$

$$f_1(4,\{2\})=8+5=13 \qquad f_1(4,\{3\})=5+5=10$$

当 $k=2$ 时,即从 1 城开始,中间经过两个城市(它们的顺序随便)到达 i 城的最短距离是

$$f_2(2,\{3,4\})=\min\left[f_1(3,\{4\})+d_{32},\quad f_1(4,\{3\})+d_{42}\right]$$

$$=\min\left[14+9,10+7\right]=17 \quad \text{所以 } p_2(2,\{3,4\})=4$$

$$f_2(3,\{2,4\})=\min\left[13+8,13+8\right]=21 \quad \text{所以 } p_2(3,\{2,4\})=2 \text{ 或 } 4$$

$$f_2(4,\{2,3\})=\min\left[14+5,16+5\right]=19 \quad \text{所以 } p_2(4,\{2,3\})=2$$

当 $k=3$ 时,即从 1 城开始,中间经过三个城市(顺序随便)回到 1 城的最短距离是

$$f_3(1,\{2,3,4\})=\min\left[f_2(2,\{3,4\})+d_{21},f_2(3,\{2,4\})+d_{31},\right.$$

$$\left.f_2(4,\{2,3\})+d_{41}\right]=\min\left[17+6,21+7,19+9\right]=23$$

所以

$$p_3(1,\{2,3,4\})=2$$

由此可知,推销员的最短旅行路线是 1→3→4→2→1,最短总距离为 23。

实际中很多问题都可以归结为货郎担这类问题。如物资运输路线中,汽车应走怎样的路线使路程最短;工厂里在钢板上要挖些小圆孔,自动焊机的割嘴应走怎样的路线使路程最短;城市里在一些地方铺设管道时,管子应走怎样的路线使管子耗费最少等等。

注　记

这一章列举了一些典型问题，介绍了如何用动态规划方法去求解。但这些问题大多数都是确定型的，而对于连续型、随机型问题接触较少。应该指出，动态规划方法是解决最优控制的一种重要方法之一，这在许多最优控制理论书中都有明确的叙述。在随机性方面，[美]R. A 霍华特著. 动态规划与马尔可夫过程. 赖炎连等译. 上海科学技术出版社，1963，以马氏过程作为模型，以动态规划的迭代方法作为手段，提供了实际计算的可能性。中国科学院应用数学研究所董泽清编著了《马尔科夫决策规划》1981年中国科学院应用数学研究所印，详细地叙述了这方面的基本理论和各种模型及应用。

至于本章所列举的典型问题，有些问题并未完全解决。如背包问题、货郎担问题，在理论和算法上还有许多问题，有些书还介绍了它们的近似算法，如卢开澄著. 组合数学——算法及分析(下册)。清华大学出版社，1983。此书就介绍了一些近似算法。排序问题在三台以上机器的情形未得到解决，中国科学院应用数学研究所越民义、韩继业写的《排序问题中的一些数学问题》(数学的实践与认识，1976年第3期和第4期)在这方面有所介绍。

习题

10.1　有一部货车每天沿着公路给四个零售店卸下6箱货物，如果各零售店出售该货物所得利润如表10-14所示，试求在各零售店卸下几箱货物，能使获得总利润最大？其值是多少？

表　10-14

箱　数	零售店			
	1	2	3	4
	利　润			
0	0	0	0	0
1	4	2	3	4
2	6	4	5	5
3	7	6	7	6
4	7	8	8	6
5	7	9	8	6
6	7	10	8	6

10.2　设有某种肥料共6个单位重量，准备供给四块粮田用。其每块田施肥数量与增产粮食数字关系如表10-15所示。试求对每块田施多少单位重量的肥料，才使总的增产粮食最多。

表 10-15

施　肥	粮　田			
	1	2	3	4
0	0	0	0	0
1	20	25	18	28
2	42	45	39	47
3	60	57	61	65
4	75	65	78	74
5	85	70	90	80
6	90	73	95	85

10.3　某公司打算向它的三个营业区增设六个销售店，每个营业区至少增设一个。从各区赚取的利润(单位为万元)与增设的销售店个数有关，其数据如下。

销售店增加数	A 区利润	B 区利润	C 区利润
0	100	200	150
1	200	210	160
2	280	220	170
3	330	225	180
4	340	230	200

试求各区应分配几个增设的销售店，才能使总利润最大？其值是多少？

10.4　某工厂有 100 台机器，拟分四个周期使用，在每一周期有两种生产任务。据经验，把机器 x_1 台投入第一种生产任务，则在一个生产周期中将有 $x_1/3$ 台机器作废；余下的机器全部投入第二种生产任务，则有 1/10 台机器作废。如果干第一种生产任务每台机器可收益 10，干第二种生产任务每台机器可收益 7。问怎样分配机器，使总收益最大？

10.5　用逐次逼近法求解下述问题

$$\max z = x_1^2 y_1 + 3x_2 y_2^2 + 4x_3^2 y_3$$

$$\begin{cases} 2x_1 + 3x_2 + 4x_3 \leqslant 24 \\ 3y_1 + 2y_2 + 5y_3 \leqslant 30 \\ x_i \geqslant 0, y_j \geqslant 0 \text{ 且为整数。} (i=1,2,3 \quad j=1,2,3) \end{cases}$$

10.6　设有三种资源，每单位的成本分别为 a、b、c。给定的利润函数为

$$r_i(x_i, y_i, z_i), \quad (i=1,2,\cdots,n)$$

现有资金为 W，应购买各种资源多少单位分配给 n 个行业，才能使总利润最大。试给出动态规划的公式，并写出它的一维递推关系式。

10.7　若例 3 每个时期的生产数量无限制，其余条件不变，试解之。

10.8　利用再生产点性质解上题。

10.9 某厂生产一种产品，估计该产品在未来4个月的销售量分别为400、500、300、200件。该项产品的生产准备费用每批为500元，每件的生产费用为1元，存储费用每件每月为1元。假定1月初的存货为100件，4月底的存货为零。试求该厂在这4个月内的最优生产计划。

10.10 某电视机厂为生产电视机而需生产喇叭，生产以万只为单位。根据以往记录，一年的四个季度需要喇叭分别是3万、2万、3万、2万只。设每万只存放在仓库内一个季度的存储费为0.2万元，每生产一批的装配费为2万元，每万只的生产成本费为1万元。问应该怎样安排四个季度的生产，才能使总的费用最小？

10.11 某公司需要对某产品决定未来半年内每个月的最佳存储量，以使总费用极小化。已知半年里对该产品的需求量和单位订货费用、单位存储费用的数据如表10-16所示。

表 10-16

月份 k	1	2	3	4	5	6
需求量 d_k	50	55	50	45	40	30
单位订货费用 c_k	825	775	850	850	775	825
单位存储费用 p_k	40	30	35	20	40	

10.12 某罐头制造公司需要在近五周内必须采购一批原料，估计在未来五周内价格有波动，其浮动价格和概率如表10-17所示。试求各周以什么价格购入，使采购价格的数学期望值最小。

表 10-17

单　价	概　率
9	0.4
8	0.3
7	0.3

10.13 求下列问题的最优解

(1) $\max z=10x_1+22x_2+17x_3$

$$\begin{cases}2x_1+4x_2+3x_3\leqslant 20\\ x_i\geqslant 0 \text{ 且为整数} \quad (i=1,2,3)\end{cases}$$

(2) $\max z=x_1x_2x_3x_4$

$$\begin{cases}2x_1+3x_2+x_3+2x_4=11\\ x_i\geqslant 0\text{。整数} \quad (i=1,2,3,4)\end{cases}$$

(3) $\max z=4x_1+5x_2+8x_3$

$$\begin{cases} x_1+x_2+x_3\leqslant 10 \\ x_1+3x_2+6x_3\leqslant 13 \\ x_i\geqslant 0,\text{整数} \quad (i=1,2,3) \end{cases}$$

(4) $\max z=g_1(x_1)+g_2(x_2)+g_3(x_3)$

$$\begin{cases} x_1^2+x_2^2+x_3^2\leqslant 20 \\ x_i\geqslant 0\ \text{且为整数} \quad (i=1,2,3) \end{cases}$$

其中 x_i 与 $g_i(x_i)$ 的关系如表 10-18 所示。

表 10-18

x_i	0	1	2	3	4	5	6	7	8	9	10
$g_1(x_1)$	2	4	7	11	13	15	18	22	18	15	11
$g_2(x_2)$	5	10	15	20	24	18	12	9	5	3	1
$g_3(x_3)$	8	12	17	22	19	16	14	11	9	7	4

10.14 某工厂生产三种产品，各产品重量与利润关系如表 10-19 所示。现将此三种产品运往市场出售，运输能力总重量不超过 6 吨。问如何安排运输使总利润最大？

表 10-19

种 类	1	2	3
重量/t	2	3	4
利润	80	130	180

10.15 某工厂在一年进行了 A、B、C 三种新产品试制，由于资金不足，估计在年内这三种新产品研制不成功的概率分别为 0.40、0.60、0.80，因而都研制不成功的概率为 $0.40\times 0.60\times 0.80=0.192$，为了促进三种新产品的研制，决定增拨 2 万元的研制费，并要资金集中使用，以万元为单位进行分配。其增拨研制费与新产品不成功的概率如表 10-20 所示。试问如何分配费用，使这三种新产品都研制不成功的概率为最小。

表 10-20

研制费 S	新产品不成功概率		
	A	B	C
0	0.40	0.60	0.80
1	0.20	0.40	0.50
2	0.15	0.20	0.30

10.16 某一印刷厂有六项加工任务，对印刷车间和装订车间所需时间(单位为天)如表 10-21 所示，试求最优的加工顺序和总加工天数。

表 10-21

车间	任务					
	J_1	J_2	J_3	J_4	J_5	J_6
印刷车间	3	10	5	2	9	11
装订车间	8	12	9	6	5	2

10.17 试制定五年中的一台机器更新策略，使总收入达到最大。设 $\alpha=1, T=2$，有关数据如表 10-22 所示。

表 10-22

项目	第一年					第二年				第三年		
	机龄											
	0	1	2	3	4	0	1	2	3	0	1	2
收入	20	19	18	16	14	25	23	22	20	27	24	22
运行费用	4	4	6	6	8	3	4	6	7	3	3	4
更新费用	25	27	30	32	35	27	29	32	34	29	30	31

项目	第四年		第五年	期前				
	机龄							
	0	1	0	2	3	4	5	6
收入	28	26	30	16	14	14	12	12
运行费用	2	3	2	6	6	7	7	8
更新费用	30	31	32	30	32	34	34	36

10.18 求解六个城市旅行推销员问题。其距离矩阵如表 10-23 所示。设推销员从 1 城出发，经过每个城市一次且仅一次，最后回到 1 城。问按怎样的路线走，使总的行程最短。

表 10-23

j	i					
	1	2	3	4	5	6
	距离					
1	0	10	20	30	40	50
2	12	0	18	30	25	21
3	23	9	0	5	10	15
4	34	32	4	0	8	16
5	45	27	11	10	0	18
6	56	22	16	20	12	0

参考资料

[1] 马仲蕃,魏权龄,赖炎连. 数学规划讲义. 北京：中国人民大学出版社,1981
[2] 俞玉森主编. 数学规划的原理和方法. 武汉：华中工学院出版社,1985
[3] 傅清祥,王晓东编著. 算法与数据结构. 北京：电子工业出版社,1998
[4] Cooper L,Cooper M W. 张有为译. 动态规划导论. 北京：国防工业出版社,1985.
[5] Bellman R E,Dreyfus S E. Applied Dynamic Programming. Princeton University Press,1962.
[6] Dreyfus S E,Law A M. The art and theory of Dynamic Programming. Academic press,1977.
[7] Sven Danø. Nonlinear and Dynamic Programming. Springer-Verlag,1975.

Part 6

第 6 篇

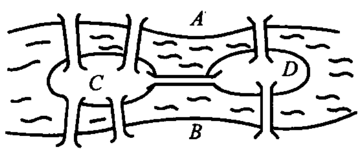
图与网络分析

CHAPTER 11
第11章

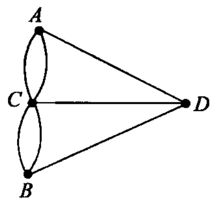
图与网络优化

图论是应用十分广泛的运筹学分支，它已广泛地应用在物理学、化学、控制论、信息论、科学管理、电子计算机等各个领域。在实际生活、生产和科学研究中，有很多问题可以用图论的理论和方法来解决。例如，在组织生产中，为完成某项生产任务，各工序之间怎样衔接，才能使生产任务完成得既快又好。一个邮递员送信，要走完他负责投递的全部街道，完成任务后回到邮局，应该按照怎样的路线走，所走的路程最短。再例如，各种通信网络的合理架设，交通网络的合理分布等问题，应用图论的方法求解都很简便。

欧拉在1736年发表图论方面的第一篇论文，解决了著名的哥尼斯堡七桥问题。哥尼斯堡城中有一条河叫普雷格尔河，该河中有两个岛，河上有七座桥。如图11-1(a)所示。



(a)



(b)

图 11-1

当时那里的居民热衷于这样的问题：一个散步者能否走过七座桥，且每座桥只走过一次，最后回到出发点。

1736年欧拉将此问题归结为如图11-1(b)所示图形的一笔画问题。即能否从某一点开始，不重复地一笔画出这个图形，最后回到出发点。欧拉证明了这是不可能的，因为图11-1(b)中的每个点都只与奇数条线相关联，不可能将这个图不重复地一笔画成。这是古典图论中的一个著名问题。

随着科学技术的发展以及电子计算机的出现与广泛应用，20世纪50年代，图论得到进一步发展。将庞大复杂的工程系统和管理问题用图描述，可以解决很多工程设计和管理决策的最优化问题。例如，完成工程任务的时间最少、距离最短、费用最省等。图论受到数学、工程技术及经营管理等各个方面越来越广泛的重视。

11.1 图的基本概念

在实际生活中,人们为了反映一些对象之间的关系,常常在纸上用点和线画出各种各样的示意图。

例 1 图 11-2 是我国北京、上海等 10 个城市间的铁路交通图,反映了这 10 个城市间的铁路分布情况。这里用点代表城市,用点和点之间的连线代表这两个城市之间的铁路线。诸如此类的还有电话线分布图、煤气管道图、航空线图等。

例 2 有甲、乙、丙、丁、戊 5 个球队,它们之间比赛的情况,也可以用图表示出来。已知甲队和其他各队都比赛过一次,乙队和甲、丙队比赛过,丙队和甲、乙、丁队比赛过,丁队和甲、丙、戊队比赛过,戊队和甲、丁队比赛过。为了反映这个情况,可以用点 v_1, v_2, v_3, v_4, v_5 分别代表这 5 个队,某两个队之间比赛过,就在这两个队所相应的点之间连一条线,这条线不过其他的点,如图 11-3 所示。

例 3 某单位储存 8 种化学药品,其中某些药品是不能存放在同一个库房里的。为了反映这个情况,可以用点 $v_1, v_2, \cdots, v_8$ 分别代表这 8 种药品,若药品 v_i 和药品 v_j 是不能存放在同一个库房的,则在 v_i 和 v_j 之间连一条线。如图 11-4 所示。从这个图中可以看到,至少要有 4 个库房,因为 v_1, v_2, v_5, v_8 必须存放在不同的库房里。事实上,4 个库房就足够了。例如,$\{v_1\}$,$\{v_2, v_4, v_7\}$,$\{v_3, v_5\}$,$\{v_6, v_8\}$各存放在一个库房里(这一类寻求库房的最少个数问题,属于图论中的所谓染色问题,一般情况下是尚未解决的)。

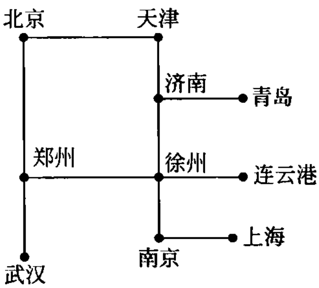


图 11-2

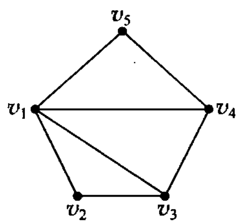


图 11-3

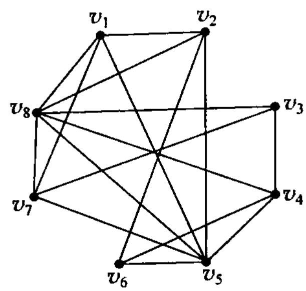


图 11-4

从以上几个例子可见,可以用由点及点与点的连线所构成的图,去反映实际生活中某些对象之间的某个特定的关系。通常用点代表研究的对象(如城市、球队、药品等),用点与点的连线表示这两个对象之间有特定的关系(如两个城市间有铁路线、两个球队比赛过、两种药品不能存放在同一个库房里等)。

因此,可以说图是反映对象之间关系的一种工具,在一般情况下,图中点的相对位置如何,点与点之间连线的长短曲直,对于反映对象之间的关系,并不重要。如例 2,也可以用图 11-5 所示的图去反映 5 个球队的比赛情况,这与图 11-3 没有本质的区别。所以,图论中的图与几何图、工程图是不同的。

前面几个例子中涉及的对象之间的“关系”具有“对称性”，就是说，如果甲与乙有这种关系，那么同时乙与甲也有这种关系。例如甲药品不能和乙药品放在一起，那么，乙药品当然也不能和甲药品放在一起。在实际生活中，有许多关系不具有这种对称性。比如人们之间的认识关系，甲认识乙并不意味着乙也认识甲。比赛中的胜负关系也是这样，甲胜乙和乙胜甲是不同的。反映这种非对称的关系，只用一条连线就不行了。如例 2，如果人们关心的是 5 个球队比赛的胜负情况，那么从图 11-3 中就看不出来了。为了反映这一类关系，可以用一条带箭头的连线表示。例如球队 v_1 胜了球队 v_2，可以从 v_1 引一条带箭头的连线到 v_2。图 11-6反映了 5 个球队比赛的胜负情况，可见 v_1 三胜一负，v_4 打了 3 场球，全负等。类似胜负这种非对称性的关系，在生产和生活中是常见的，如交通运输中的“单行线”，部门之间的领导与被领导的关系、一项工程中各工序之间的先后关系等。

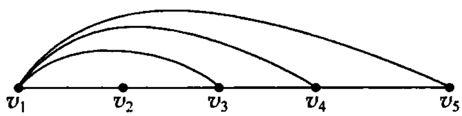


图　11-5

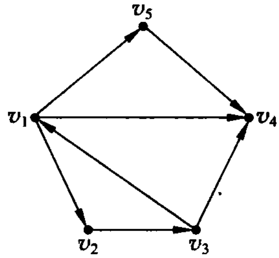


图　11-6

综上所述，一个图是由一些点及一些点之间的连线（不带箭头或带箭头）所组成的。

为了区别起见，把两点之间的不带箭头的连线称为**边**，带箭头的连线称为**弧**。

如果一个图 G 是由点及边所构成的，则称之为**无向图**（也简称为**图**），记为 $G=(V,E)$，式中 V,E 分别是 G 的点集合和边集合。一条连结点 $v_i,v_j\in V$ 的边记为 $[v_i,v_j]$（或 $[v_j,v_i]$）。

如果一个图 D 是由点及弧所构成的，则称为**有向图**，记为 $D=(V,A)$，式中 V,A 分别表示 D 的点集合和弧集合。一条方向是从 v_i 指向 v_j 的弧记为 (v_i,v_j)。

图 11-7 是一个无向图。

$$V=\{v_1,v_2,v_3,v_4\},E=\{e_1,e_2,e_3,e_4,e_5,e_6,e_7\}$$

其中

$$e_1=[v_1,v_2],e_2=[v_1,v_2],e_3=[v_2,v_3],e_4=[v_3,v_4]$$
$$e_5=[v_1,v_4],e_6=[v_1,v_3],e_7=[v_4,v_4]$$

图 11-8 是一个有向图。

$$V=\{v_1,v_2,v_3,v_4,v_5,v_6,v_7\},A=\{a_1,a_2,a_3,a_4,\cdots,a_{11}\}$$

其中

$$a_1=(v_1,v_2),a_2=(v_1,v_3),a_3=(v_3,v_2),a_4=(v_3,v_4)$$
$$a_5=(v_2,v_4),a_6=(v_4,v_5),a_7=(v_4,v_6),a_8=(v_5,v_3)$$

$a_9=(v_5,v_4),a_{10}=(v_5,v_6),a_{11}=(v_6,v_7)$

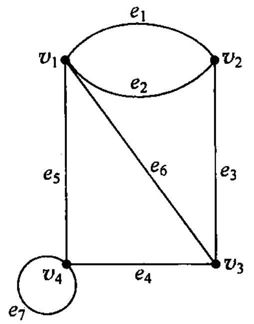


图 11-7

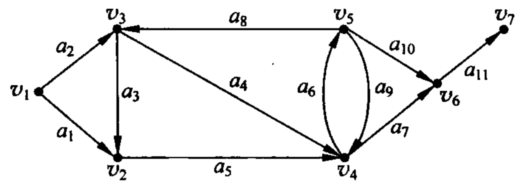


图 11-8

图 G 或 D 中的点数记为 $p(G)$ 或 $p(D)$，边(弧)数记为 $q(G)(q(D))$。在不会引起混淆的情况下，也分别简记为 p,q。

下面介绍常用的一些名词和记号，先考虑无向图 $G=(V,E)$。

若边 $e=[u,v]\in E$，则称 u,v 是 e 的**端点**，也称 u,v 是**相邻的**。称 e 是点 u(及点 v)的**关联边**。若图 G 中，某个边 e 的两个端点相同，则称 e 是**环**(如图 11-7 中的 e_7)，若两个点之间有多于一条的边，称这些边为**多重边**(如图 11-7 中的 e_1,e_2)。一个无环，无多重边的图称为**简单图**，一个无环，但允许有多重边的图称为**多重图**。

以点 v 为端点的边的个数称为 v 的**次**，记为 $d_G(v)$ 或 $d(v)$。如图 11-7 中，$d(v_1)=4$，$d(v_2)=3,d(v_3)=3,d(v_4)=4$(环 e_7 在计算 $d(v_4)$ 时算作两次)。

称次为 1 的点为**悬挂点**，悬挂点的关联边称为**悬挂边**，次为零的点称为**孤立点**。

定理 1 图 $G=(V,E)$ 中，所有点的次之和是边数的两倍，即

$$\sum_{v\in V}d(v)=2q$$

这是显然的，因为在计算各点的次时，每条边被它的端点各用了一次。

次为奇数的点，称为**奇点**，否则称为**偶点**。

定理 2 任一个图中，奇点的个数为偶数。

证明 设 V_1 和 V_2 分别是 G 中奇点和偶点的集合，由定理 1，有

$$\sum_{v\in V_1}d(v)+\sum_{v\in V_2}d(v)=\sum_{v\in V}d(v)=2q$$

因 $\sum\limits_{v\in V}d(v)$ 是偶数，$\sum\limits_{v\in V_2}d(v)$ 也是偶数，故 $\sum\limits_{v\in V_1}d(v)$ 必定也是偶数，从而 V_1 的点数是偶数。

给定一个图 $G=(V,E)$，一个点、边的交错序列 $(v_{i_1},e_{i_1},v_{i_2},e_{i_2},\cdots,v_{i_{k-1}},e_{i_{k-1}},v_{i_k})$，如果满足 $e_{i_t}=[v_{i_t},v_{i_{t+1}}]\ (t=1,2,\cdots,k-1)$，则称为一条联结 v_{i_1} 和 v_{i_k} 的**链**，记为 $(v_{i_1},v_{i_2},\cdots,v_{i_k})$，有时称点 $v_{i_2},v_{i_3},\cdots,v_{i_{k-1}}$ 为链的**中间点**。

链 $(v_{i_1},v_{i_2},\cdots,v_{i_k})$ 中，若 $v_{i_1}=v_{i_k}$，则称之为一个**圈**，记为 $(v_{i_1},v_{i_2},\cdots,v_{i_{k-1}},v_{i_1})$。若

链$(v_{i_1},v_{i_2},\cdots,v_{i_k})$中，点 $v_{i_1},v_{i_2},\cdots,v_{i_k}$ 都是不同的，则称之为**初等链**；若圈$(v_{i_1},v_{i_2},\cdots,v_{i_{k-1}},v_{i_1})$中，$v_{i_1},v_{i_2},\cdots,v_{i_{k-1}}$ 都是不同的，则称之为**初等圈**；若链(圈)中含的边均不相同，则称之为**简单链(圈)**。以后说到链(圈)，除非特别交代，均指初等链(圈)。

例如图 11-9 中，$(v_1,v_2,v_3,v_4,v_5,v_3,v_6,v_7)$是一条简单链，但不是初等链，$(v_1,v_2,v_3,v_6,v_7)$是一条初等链。这个图中，不存在联结 v_1 和 v_9 的链。(v_1,v_2,v_3,v_4,v_1)是一个初等圈，$(v_4,v_1,v_2,v_3,v_5,v_7,v_6,v_3,v_4)$是简单圈，但不是初等圈。

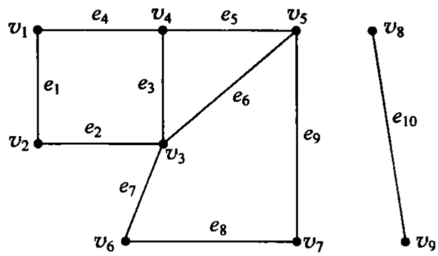


图 11-9

图 G 中，若任何两个点之间，至少有一条链，则称 G 是**连通图**，否则称为**不连通图**。若 G 是不连通图，它的每个连通的部分称为 G 的一个**连通分图**(也简称**分图**)。如图 11-9 是一个不连通图，它有两个连通分图。

给了一个图 $G=(V,E)$，如果图 $G'=(V',E')$，使 $V=V'$ 及 $E'\subseteq E$，则称 G' 是 G 的一个**支撑子图**。

设 $v\in V(G)$，用 G-v 表示从图 G 中去掉点 v 及 v 的关联边后得到的一个图。

例如若 G 如图 11-10(a)所示，则 G-v_3 见图 11-10(b)。图 11-10(c)是图 G 的一个支撑子图。

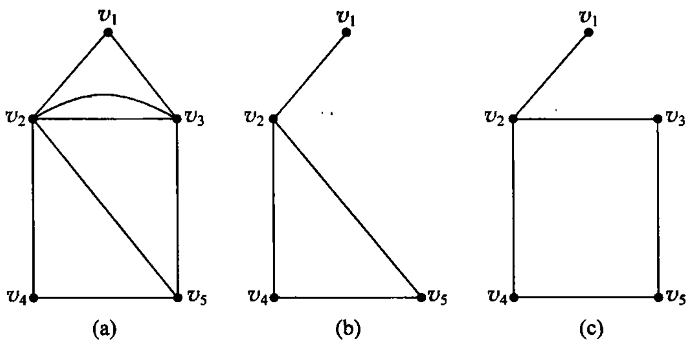


图 11-10

现在讨论有向图的情形。设给了一个有向图，$D=(V,A)$，从 D 中去掉所有弧上的箭头，就得到一个无向图，称之为 D 的**基础图**，记之为 $G(D)$。

给 D 中的一条弧 $a=(u,v)$，称 u 为 a 的**始点**，v 为 a 的**终点**，称弧 a 是从 u 指向 v 的。

设$(v_{i_1},a_{i_1},v_{i_2},a_{i_2},\cdots,v_{i_{k-1}},a_{i_{k-1}},v_{i_k})$是 D 中的一个点弧交错序列，如果这个序列在基础图 $G(D)$ 中所对应的点边序列是一条链，则称这个点弧交错序列是 D 的一条链。类似定义圈和初等链(圈)。

如果$(v_{i_1},a_{i_1},v_{i_2},a_{i_2},\cdots,v_{i_{k-1}},a_{i_{k-1}},v_{i_k})$是 D 中的一条链，并且对 $t=1,2,\cdots,k-1$，

均有 $a_{i_t}=(v_{i_t},v_{i_{t+1}})$，称之为**从 v_{i_1} 到 v_{i_k} 的一条路**。若路的第一个点和最后一点相同，则称之为**回路**。类似定义**初等路**(回路)。

例如图 11-8 中，$(v_3,(v_3,v_2),v_2,(v_2,v_4),v_4,(v_4,v_5),v_5,(v_5,v_3),v_3)$是一个回路，$(v_1,(v_1,v_3),v_3,(v_3,v_4),v_4,(v_4,v_6),v_6)$是从 v_1 到 v_6 的路，$(v_1,(v_1,v_3),v_3,(v_5,v_3),v_5,(v_5,v_6),v_6)$是一条链，但不是路。

对无向图，链与路(圈与回路)这两个概念是一致的。

类似于无向图，可定义简单有向图、多重有向图，图 11-8 是一个简单有向图。以后除特别交代外，说到图(有向图)均指简单图(简单有向图)。

11.2 树

11.2.1 树及其性质

在各式各样的图中，有一类图是极其简单然而却是很有用的，这就是树。

例 4 已知有五个城市，要在它们之间架设电话线，要求任何两个城市都可以互相通话(允许通过其他城市)，并且电话线的根数最少。

用五个点 v_1,v_2,v_3,v_4,v_5 代表五个城市，如果在某两个城市之间架设电话线，则在相应的两个点之间连一条边，这样一个电话线网就可以用一个图来表示。首先，为了使任何两个城市都可以通话，这样的图必须是连通的。其次，若图中有圈的话，从圈上任意去掉一条边，余下的图仍是连通的，这样可以省去一根电话线。因而，满足要求的电话线网所对应的图必定是不含圈的连通图。图 11-11 代表了满足要求的一个电话线网。

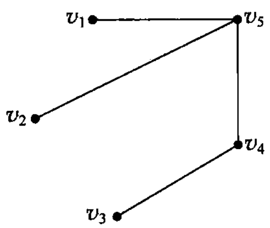


图 11-11

定义 1 一个无圈的连通图称为**树**。

例 5 某工厂的组织机构如图 11-12 所示。如果用图表示，该工厂的组织机构图就是一个树(如图 11-13 所示)。

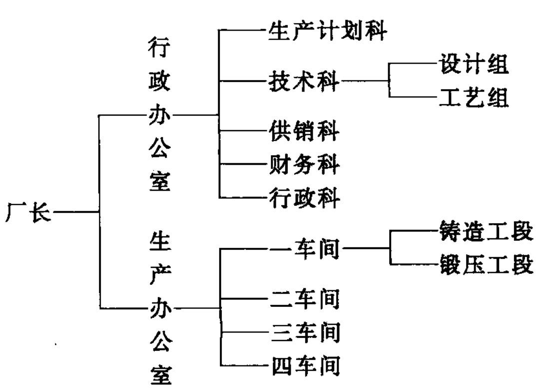
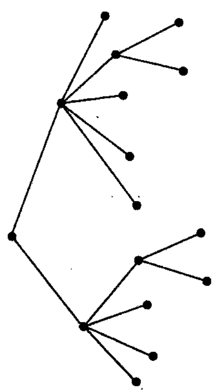


图 11-12


图 11-13

下面介绍树的一些重要性质。

定理 3 设图 $G=(V,E)$ 是一个树，$p(G)\geqslant 2$，则 G 中至少有两个悬挂点。

证明 令 $P=(v_1,v_2,\cdots,v_k)$ 是 G 中含边数最多的一条初等链，因 $p(G)\geqslant 2$，并且 G 是连通的，故链 P 中至少有一条边，从而 v_1 与 v_k 是不同的。现在来证明：v_1 是悬挂点，即 $d(v_1)=1$。用反证法，如果 $d(v_1)\geqslant 2$，则存在边 $[v_1,v_m]$，使 $m\neq 2$。若点 v_m 不在 P 上，那么 $(v_m,v_1,v_2,\cdots,v_k)$ 是 G 中的一条初等链，它含的边数比 P 多一条，这与 P 是含边数最多的初等链矛盾。若点 v_m 在 P 上，那么 $(v_1,v_2,\cdots,v_m,v_1)$ 是 G 中的一个圈，这与树的定义矛盾。于是必有 $d(v_1)=1$，即 v_1 是悬挂点。同理可证 v_k 也是悬挂点，因而 G 至少有两个悬挂点。

定理 4 图 $G=(V,E)$ 是一个树的充分必要条件是 G 不含圈，且恰有 $p-1$ 条边。

证明 必要性 设 G 是一个树，根据定义，G 不含圈，故只要证明 G 恰有 $p-1$ 条边。对点数 p 施行数学归纳法。$p=1,2$ 时，结论显然成立。

假设对点数 $p\leqslant n$ 时，结论成立。设树 G 含 $n+1$ 个点。由定理 3，G 含悬挂点，设 v_1 是 G 的一个悬挂点，考虑图 $G-v_1$，易见 $p(G-v_1)=n$，$q(G-v_1)=q(G)-1$。因 $G-v_1$ 是 n 个点的树，由归纳假设，$q(G-v_1)=n-1$，于是

$$q(G)=q(G-v_1)+1=(n-1)+1=n=p(G)-1$$

充分性 只要证明 G 是连通的。用反证法，设 G 是不连通的，G 含 s 个连通分图 $G_1,G_2,\cdots,G_s(s\geqslant 2)$。因每个 $G_i(i=1,2,\cdots,s)$ 是连通的，并且不含圈，故每个 G_i 是树。设 G_i 有 p_i 个点，则由必要性，G_i 有 p_i-1 条边，于是

$$q(G)=\sum_{i=1}^{s}q(G_i)=\sum_{i=1}^{s}(p_i-1)=\sum_{i=1}^{s}p_i-s=p(G)-s\leqslant p(G)-2$$

这与 $q(G)=p(G)-1$ 的假设矛盾。

定理 5 图 $G=(V,E)$ 是一个树的充分必要条件是 G 是连通图，并且

$$q(G)=p(G)-1$$

证明 必要性 设 G 是树，根据定义，G 是连通图，由定理 4，$q(G)=p(G)-1$

充分性 只要证明 G 不含圈，对点数施行归纳。$p(G)=1,2$ 时，结论显然成立。设 $p(G)=n(n\geqslant 1)$ 时结论成立。现设 $p(G)=n+1$，首先证明 G 必有悬挂点。若不然，因 G 是连通的，且 $p(G)\geqslant 2$，故对每个点 v_i，有 $d(v_i)\geqslant 2$。从而

$$q(G)=\frac{1}{2}\sum_{i=1}^{p(G)}d(v_i)\geqslant p(G)$$

这与 $q(G)=p(G)-1$ 矛盾，故 G 必有悬挂点。设 v_1 是 G 的一个悬挂点，考虑 $G-v_1$，这个图仍是连通的，$q(G-v_1)=q(G)-1=p(G)-2=p(G-v_1)-1$，由归纳假设知 $G-v_1$ 不含圈，于是 G 也不含圈。

定理 6 图 G 是树的充分必要条件是任意两个顶点之间恰有一条链。

证明 必要性 因 G 是连通的，故任两个点之间至少有一条链。但如果某两个点之

间有两条链的话，那么图 G 中含有圈，这与树的定义矛盾，从而任两个点之间恰有一条链。

充分性　设图 G 中任两个点之间恰有一条链，那么易见 G 是连通的。如果 G 中含有圈，那么这个圈上的两个顶点之间有两条链，这与假设矛盾，故 G 不含圈，于是 G 是树。

由这个定理，很容易推出如下结论：

(1) 从一个树中去掉任意一条边，则余下的图是不连通的。由此可知，在点集合相同的所有图中，树是含边数最少的连通图。这样，例 4 中所要求的电话线网就是以这五个城市为点的一个树。

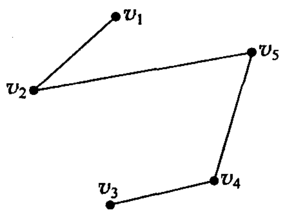


图　11-14

(2) 在树中不相邻的两个点间添上一条边，则恰好得到一个圈。如果再从这个圈上任意去掉一条边，可以得到一个树。

如图 11-11 中，添加 $[v_2, v_1]$，就得到一个圈 (v_1, v_2, v_5, v_1)，如果从这个圈中去掉一条边 $[v_1, v_5]$，就得到如图 11-14 所示的树。

11.2.2　图的支撑树

定义 2　设图 $T=(V,E')$ 是图 $G=(V,E)$ 的支撑子图，如果图 $T=(V,E')$ 是一个树，则称 T 是 G 的一个**支撑树**。

例如图 11-15(b) 是图 11-15(a) 所示图的一个支撑树。

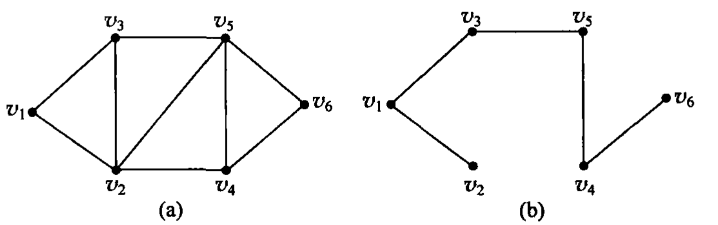


图　11-15

若 $T=(V,E')$ 是 $G=(V,E)$ 的一个支撑树，则显然，树 T 中边的个数是 $p(G)-1$，G 中不属于树 T 的边数是 $q(G)-p(G)+1$

定理 7　图 G 有支撑树的充分必要条件是图 G 是连通的。

证明　必要性是显然的。

充分性　设图 G 是连通图，如果 G 不含圈，那么 G 本身是一个树，从而 G 是它自身的一个支撑树。现设 G 含圈，任取一个圈，从圈中任意地去掉一条边，得到图 G 的一个支撑子图 G_1。如果 G_1 不含圈，那么 G_1 是 G 的一个支撑树(因为易见 G_1 是连通的)；如果 G_1 仍含圈，那么从 G_1 中任取一个圈，从圈中再任意去掉一条边，得到图 G 的一个支撑子图

G_2，如此重复，最终可以得到 G 的一个支撑子图 G_k，它不含圈，于是 G_k 是 G 的一个支撑树。

定理 5 充分性的证明，提供了一个寻求连通图的支撑树的方法。这就是任取一个圈，从圈中去掉一边，对余下的图重复这个步骤，直到不含圈时为止，即得到一个支撑树，称这种方法为"**破圈法**"。

例 6 在图 11-16 中，用破圈法求出图的一个支撑树。

解 取一个圈(v_1,v_2,v_3,v_1)，从这个圈中去掉边 $e_3=[v_2,v_3]$；在余下的图中，再取一个圈(v_1,v_2,v_4,v_3,v_1)，去掉边 $e_4=[v_2,v_4]$；在余下的图中，从圈(v_3,v_4,v_5,v_3)中去掉边 $e_6=[v_5,v_3]$；再从圈$(v_1,v_2,v_5,v_4,v_3,v_1)$中去掉边 $e_8=[v_2,v_5]$。这时，剩余的图中不含圈，于是得到一个支撑树，如图 11-16 中粗线所示。

也可以用另一种方法来寻求连通图的支撑树。在图中任取一条边 e_1，找一条与 e_1 不构成圈的边 e_2，再找一条与$\{e_1,e_2\}$不构成圈的边 e_3，一般，设已有$\{e_1,e_2,\cdots,e_k\}$，找一条与$\{e_1,e_2,\cdots,e_k\}$中的任何一些边不构成圈的边 e_{k+1}。重复这个过程，直到不能进行为止。这时，由所有取出的边所构成的图是一个支撑树，称这种方法为"**避圈法**"。

例 7 在图 11-17 中，用避圈法求出一个支撑树。

解 首先任取边 e_1，因 e_2 与 e_1 不构成圈，所以可以取 e_2，因为 e_5 与$\{e_1,e_2\}$不构成圈，故可以取 e_5（因 e_3 与$\{e_1,e_2\}$构成一个圈(v_1,v_2,v_3,v_1)，所以不能取 e_3）；因 e_6 与$\{e_1,e_2,e_5\}$不构成圈，故可取 e_6；因 e_8 与$\{e_1,e_2,e_5,e_6\}$不构成圈，故可取 e_8（注意，因 e_7 与$\{e_1,e_2,e_5,e_6\}$中的 e_5,e_6 构成圈(v_2,v_5,v_4,v_2)，故不能取 e_7）。这时由$\{e_1,e_2,e_5,e_6,e_8\}$所构成的图就是一个支撑树，如图 11-17 中粗线所示。

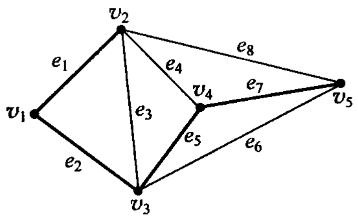


图 11-16

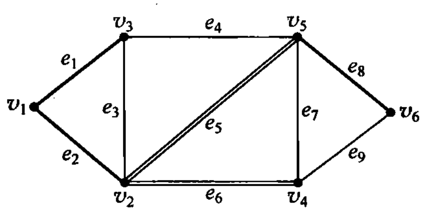


图 11-17

实际上，由定理 4，5 可知，在"破圈法"中去掉的边数必是 $q(G)-p(G)+1$ 条，在"避圈法"中取出的边数必定是 $p(G)-1$ 条。

11.2.3 最小支撑树问题

定义 3 给图 $G=(V,E)$，对 G 中的每一条边$[v_i,v_j]$，相应地有一个数 w_{ij}，则称这样的图 G 为**赋权图**，w_{ij} 称为边$[v_i,v_j]$上的**权**。

这里所说的"权"，是指与边有关的数量指标。根据实际问题的需要，可以赋予它不同的含义，例如表示距离、时间、费用等。

赋权图在图的理论及其应用方面有着重要的地位。赋权图不仅指出各个点之间的邻接关系,而且同时也表示出各点之间的数量关系。所以,赋权图被广泛地应用于解决工程技术及科学生产管理等领域的最优化问题。最小支撑树问题就是赋权图上的最优化问题之一。

设有一个连通图 $G=(V,E)$,每一边 $e=[v_i,v_j]$,有一个非负权

$$w(e)=w_{ij}\quad(w_{ij}\geqslant 0)$$

定义 4 如果 $T=(V,E')$ 是 G 的一个支撑树,称 E' 中所有边的权之和为支撑树 T 的权,记为 $w(T)$。即

$$w(T)=\sum_{[v_i,v_j]\in T} w_{ij}$$

如果支撑树 T^* 的权 $w(T^*)$ 是 G 的所有支撑树的权中最小者,则称 T^* 是 G 的**最小支撑树(简称最小树)**。即

$$w(T^*)=\min_T w(T)$$

式中对 G 的所有支撑树 T 取最小。

最小支撑树问题就是要求给定连通赋权图 G 的最小支撑树。

假设给定一些城市,已知每对城市间交通线的建造费用。要求建造一个联结这些城市的交通网,使总的建造费用最小,这个问题就是赋权图上的最小树问题。

下面介绍求最小树的两个方法。

1. 避圈法(Kruskal)

开始选一条最小权的边,以后每一步中,总从与已选边不构成圈的那些未选边中,选一条权最小的。(每一步中,如果有两条或两条以上的边都是权最小的边,则从中任选一条。)

算法的具体步骤如下:给定赋权图 $G=(V,E)$

第一步:令 $i=1,E_0=\varnothing$,($\varnothing$表示空集)。

第二步:

(2.1)如果 $i=p(G)$,那么 $T=(V,E_{i-1})$ 是最小支撑树,算法终止。

(2.2) 如果 $i<p(G)$,选一条边 $e_i\in E\backslash E_{i-1}$,使 e_i 是使 $(V,E_{i-1}\cup\{e\})$ 不含圈的所有边 $e(e\in E\backslash E_{i-1})$ 中权最小的边。如果这样的边不存在,则说明图 G 不含支撑树,从而也就没有最小支撑树,算法终止。否则,令 $E_i=E_{i-1}\cup\{e_i\}$,转入第三步。

第三步:把 i 换成 $i+1$,转入第二步。

在证明这个方法的正确性之前,先介绍一个例子。

例 8 某工厂内联结六个车间的道路网如图 11-18(a)所示。已知每条道路的长,要求沿道路架设联结六个车间的电话线网,使电话线的总长最小。

解 这个问题就是求如图 11-18(a)所示的赋权图上的最小树,用避圈法求解。

$i=1,E_0=\varnothing$,从 E 中选最小权边 $[v_2,v_3]$,$E_1=\{[v_2,v_3]\}$;

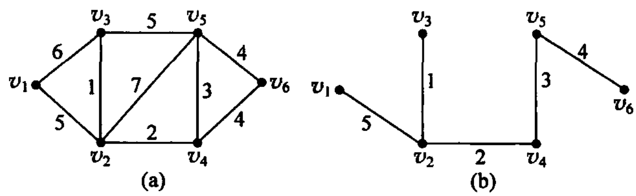


图 11-18

$i=2$,从 $E\backslash E_1$ 中选最小权边$[v_2,v_4]$($[v_2,v_4]$与$[v_2,v_3]$不构成圈),$E_2=\{[v_2,v_3],[v_2,v_4]\}$;

$i=3$,从 $E\backslash E_2$ 中选$[v_4,v_5]$($(V,E_2\cup\{[v_4,v_5]\})$不含圈),令 $E_3=\{[v_2,v_3],[v_2,v_4],[v_4,v_5]\}$;

$i=4$,从 $E\backslash E_3$ 中选$[v_5,v_6]$(或选$[v_4,v_6]$)($(V,E_3\cup\{[v_5,v_6]\})$不含圈),令 $E_4=\{[v_2,v_3],[v_2,v_4],[v_4,v_5],[v_5,v_6]\}$;

$i=5$,从 $E\backslash E_4$ 中选$[v_1,v_2]$($(V,E_4\cup\{[v_1,v_2]\})$不含圈)。注意,因$[v_4,v_6]$与已选边$[v_4,v_5]$,$[v_5,v_6]$构成圈,所以虽然$[v_4,v_6]$的权小于$[v_1,v_2]$的权,但这时不能选$[v_4,v_6]$,令 $E_5=\{[v_2,v_3],[v_2,v_4],[v_4,v_5],[v_5,v_6],[v_1,v_2]\}$;

$i=6$,这时,任一条未选的边都与已选的边构成圈,所以算法终止。

(V,E_5)就是要求的最小树,即电话线总长最小的电话线网方案(图 11-18(b)),电话线总长为 15 单位。

现在来证明方法一的正确性。

令 $G=(V,E)$是连通赋权图,我们只要证明:$i=p(G)$时,$T=(V,E_{i-1})$是最小支撑树。

首先根据定理 4,可知 $T=(V,E_{i-1})$是支撑树。为方便起见,记

$$E(T)=\{e_1,e_2,\cdots,e_{p-1}\}$$

式中,$p=p(G)$。

用反证法来证明 T 是最小支撑树,设 T 不是最小支撑树,在 G 的所有支撑树中,令 H 是与 T 的公共边数最大的最小支撑树。因 T 与 H 不是同一个支撑树,故 T 中至少有一条边不在 H 中。令 $e_i(1\leqslant i\leqslant p-1)$是第一个不属于 H 的边,把 e_i 放入 H 中,必得到一个且仅一个圈,记这个圈为 C。因为 T 是不含圈的,故 C 中必有一条边不属于 T,记这条边为 e。在 H 中去掉 e,增加 e_i,就得到 G 的另一个支撑树 T_0,可见

$$w(T_0)=w(H)+w(e_i)-w(e)$$

因为 $w(H)\leqslant w(T_0)$(因 H 是最小支撑树),推出 $w(e)\leqslant w(e_i)$。但根据算法,e_i 是使$(V,\{e_1,e_2,\cdots,e_i\})$不含圈的权最小的边,而$(V,\{e_1,e_2,\cdots,e_{i-1},e\})$也是不含圈的,故必有 $w(e)=w(e_i)$,从而 $w(T_0)=w(H)$。这就是说 T_0 也是 G 的一个最小支撑树,但是 T_0 与 T 的公共边数比 H 与 T 的公共边数多一条,这与 H 的选取矛盾。

2. 破圈法

任取一个圈,从圈中去掉一条权最大的边(如果有两条或两条以上的边都是权最大的边,则任意去掉其中一条)。在余下的图中,重复这个步骤,直至得到一个不含圈的图为止,这时的图便是最小树。

例 9 用破圈法求图 11-18(a)所示赋权图的最小支撑树。

解 任取一个圈,比如(v_1,v_2,v_3,v_1),边$[v_1,v_3]$是这个圈中权最大的边,于是丢去$[v_1,v_3]$;再取圈(v_3,v_5,v_2,v_3),去掉$[v_2,v_5]$;取圈(v_3,v_5,v_4,v_2,v_3),去掉$[v_3,v_2]$;取圈(v_5,v_6,v_4,v_5),这个圈中,$[v_5,v_6]$及$[v_4,v_6]$都是权最大的边,去掉其中的一条,比如说$[v_4,v_6]$。这时得到一个不含圈的图(如图 11-18(b)所示),即为最小树。

关于破圈法的正确性略去证明。

11.3 最短路问题

11.3.1 引例

例 10 已知如图 11-19 所示的单行线交通网,每弧旁的数字表示通过这条单行线所需要的费用。现在某人要从 v_1 出发,通过这个交通网到 v_8 去,求使总费用最小的旅行路线。

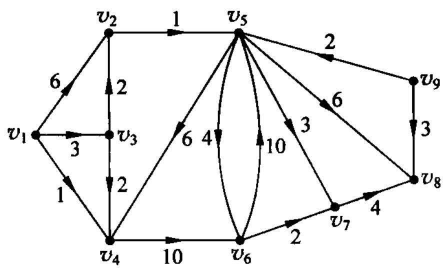


图 11-19

可见,从 v_1 到 v_8 的旅行路线是很多的,例如可以从 v_1 出发,依次经过 v_2,v_5,然后到 v_8;也可以从 v_1 出发,依次经过 v_3,v_4,v_6,v_7,然后到 v_8 等。不同的路线,所需总费用是不同的。比如,按前一个路线,总费用是 6+1+6=13 单位;而按后一个路线,总费用是 3+2+10+2+4=21 单位。不难看到,用图的语言来描述,从 v_1 到 v_8 的旅行路线与有向图中从 v_1 到 v_8 的路是一一对应的。一条旅行路线的总费用就是相应的从 v_1 到 v_8 的路中所有弧旁数字之和。当然,这里说到的路可以不是初等路。例如某人从 v_1 到 v_8 的旅行路线可以是从 v_1 出发,依次经 $v_3,v_4,v_6,v_5,v_4,v_6,v_7$,最后到达 v_8。这条路线相应的路是$(v_1,v_3,v_4,v_6,v_5,v_4,v_6,v_7,v_8)$,总费用是 47 单位。

从这个例子可以引出一般的最短路问题,给定一个赋权有向图,即给了一个有向图 $D=(V,A)$,对每一个弧 $a=(v_i,v_j)$,相应地有权 $w(a)=w_{ij}$,又给定 D 中的两个顶点 v_s,v_t。设 P 是 D 中从 v_s 到 v_t 的一条路,定义路 P 的权是 P 中所有弧的权之和,记为 $w(P)$。最短路问题就是要在所有从 v_s 到 v_t 的路中,求一条权最小的路,即求一条从 v_s 到 v_t 的路 P_0,使

$$w(P_0)=\min_{P} w(P)$$

式中对 D 中所有从 v_s 到 v_t 的路 P 取最小，称 P_0 是从 v_s 到 v_t 的**最短路**。路 P_0 的权称为从 v_s 到 v_t 的**距离**，记为 $d(v_s,v_t)$。显然，$d(v_s,v_t)$与 $d(v_t,v_s)$不一定相等。

最短路问题是重要的最优化问题之一，它不仅可以直接应用于解决生产实际的许多问题，如管道铺设、线路安排、厂区布局、设备更新等，而且经常被作为一个基本工具，用于解决其他的优化问题。

11.3.2 最短路算法

本节将介绍在一个赋权有向图中寻求最短路的方法，这些方法实际上求出了从给定一个点 v_s 到任一个点 v_j 的最短路。

如下事实是经常要利用的，如果 P 是 D 中从 v_s 到 v_j 的最短路，v_i 是 P 中的一个点，那么，从 v_s 沿 P 到 v_i 的路是从 v_s 到 v_i 的最短路。事实上，如果这个结论不成立，设 Q 是从 v_s 到 v_i 的最短路，令 P'是从 v_s 沿 Q 到达 v_i，再从 v_i 沿 P 到达 v_j 的路，那么，P'的权就比 P 的权小，这与 P 是从 v_s 到 v_j 的最短路矛盾。

首先介绍所有 $w_{ij}\geqslant 0$ 的情形下，求最短路的方法。当所有的 $w_{ij}\geqslant 0$ 时，目前公认最好的方法是由 Dijkstra 于 1959 年提出来的。

Dijkstra 方法的基本思想是从 v_s 出发，逐步地向外探寻最短路。执行过程中，与每个点对应，记录下一个数(称为这个点的标号)，它或者表示从 v_s 到该点的最短路的权(称为 P 标号)，或者是从 v_s 到该点的最短路的权的上界(称为 T 标号)，方法的每一步是去修改 T 标号，并且把某一个具 T 标号的点改变为具 P 标号的点，从而使 D 中具 P 标号的顶点数多一个，这样，至多经过 $p-1$ 步，就可以求出从 v_s 到各点的最短路。

在叙述 Dijkstra 方法的具体步骤之前，以例 10 为例说明一下这个方法的基本思想。例 10 中，$s=1$。因为所有 $w_{ij}\geqslant 0$，故有 $d(v_1,v_1)=0$。这时，v_1 是具 P 标号的点。现在考查从 v_1 发出的三条弧，(v_1,v_2)，(v_1,v_3)和(v_1,v_4)。如果某人从 v_1 出发沿(v_1,v_2)到达 v_2，这时需要 $d(v_1,v_1)+w_{12}=6$ 单位的费用；如果他从 v_1 出发，沿(v_1,v_3)到达 v_3，则需要 $d(v_1,v_1)+w_{13}=3$ 单位的费用；类似地，若沿(v_1,v_4)到达 v_4，需要 $d(v_1,v_1)+w_{14}=1$ 单位的费用。因

$$\min\{d(v_1,v_1)+w_{12},d(v_1,v_1)+w_{13},d(v_1,v_1)+w_{14}\}=d(v_1,v_1)+w_{14}=1$$

可以断言，他从 v_1 出发到 v_4 所需要的最小费用必定是 1 单位，即从 v_1 到 v_4 的最短路是(v_1,v_4)，$d(v_1,v_4)=1$。这是因为从 v_1 到 v_4 的任一条路 P，如果不是(v_1,v_4)，则必是先从 v_1 沿(v_1,v_2)到达 v_2，或者沿(v_1,v_3)到达 v_3，而后再从 v_2 或 v_3 到 v_4 去，但如上所说，这时候他已需要 6 单位或 3 单位的费用，不管他如何再从 v_2 或从 v_3 到达 v_4，所需要的总费用都不会比 1 少(因为所有的 $w_{ij}\geqslant 0$)。因而推知 $d(v_1,v_4)=1$，这样就可以使 v_4 变成具 P 标号的点。

现在考查从 v_1 及 v_4 指向其余点的弧，由上已知，从 v_1 出发，分别沿(v_1,v_2)、(v_1,v_3)到达(v_2,v_3)，需要 6 单位或 3 单位的费用，而从 v_4 出发沿(v_4,v_6)到达 v_6，所需要的费用

是 $d(v_1,v_4)+w_{46}=1+10=11$ 单位，因

$$\min\{d(v_1,v_1)+w_{12},d(v_1,v_1)+w_{13},d(v_1,v_4)+w_{46}\}=d(v_1,v_1)+w_{13}=3$$

基于同样的理由可以断言，从 v_1 到 v_3 的最短路是(v_1,v_3)，$d(v_1,v_3)=3$。这样又可以使点 v_3 变成具 P 标号的点。如此重复这个过程，可以求出从 v_1 到任一点的最短路。

在下述 Dijkstra 方法具体步骤中，用 $P(v)$，$T(v)$分别表示点 v 的 P 标号和 T 标号，S_i 表示第 i 步时，具 P 标号的点的集合。为了在求出从 v_s 到各点的距离的同时，也求出从 v_s 到各点的最短路，给每个点 v 以一个 λ 值，算法终止时，如果 $\lambda(v)=m$，表示在从 v_s 到 v 的最短路上，v 的前一个点是 v_m；如果 $\lambda(v)=M$，则表示 D 中不含从 v_s 到 v 的路；$\lambda(v)=0$ 表示 $v=v_s$。

Dijkstra 方法的具体步骤：给定赋权有向图 $D=(V,A)$。

开始$(i=1)$令 $S_1=\{v_s\}$，$P(v_s)=0$，$\lambda(v_s)=0$，对每一个 $v\neq v_s$，令 $T(v)=+\infty$，$\lambda(v)=M$，令 $k=s$。

第一步：如果 $S_i=V$，算法终止，这时，对每个 $v\in S_i$，$d(v_s,v)=P(v)$；否则转入第二步。

第二步：考查每个使$(v_k,v_j)\in A$ 且 $v_j\notin S_i$ 的点 v_j。如果 $T(v_j)>P(v_k)+w_{kj}$，则把 $T(v_j)$修改为 $P(v_k)+w_{kj}$，把 $\lambda(v_j)$修改为 k；否则转入第 3 步。

第三步：令 $T(v_{j_i})=\min\limits_{v_j\notin S_i}\{T(v_j)\}$。

如果 $T(v_{j_i})<+\infty$，则把 v_{j_i} 的 T 标号变为 P 标号 $P(v_{j_i})=T(v_{j_i})$，令 $S_{i+1}=S_i\cup\{v_{j_i}\}$，$k=j_i$，把 i 换成 $i+1$，转入第一步；否则算法终止，这时对每一个 $v\in S_i$，$d(v_s,v)=P(v)$，而对每一个 $v\notin S_i$，$d(v_s,v)=T(v)$。

现在用 Dijkstra 方法求例 10 中从 v_1 到各个顶点的最短路，这时 $s=1$。

(1) $i=1$

$S_1=\{v_1\}$，$P(v_1)=0$，$\lambda(v_1)=0$，$T(v_i)=+\infty$，$\lambda(v_i)=M(i=2,3,\cdots,9)$，以及$k=1$。

转入第二步，因$(v_1,v_2)\in A$，$v_1\notin S_1$，$P(v_1)+w_{12}<T(v_2)$，故把 $T(v_2)$修改为 $P(v_1)+w_{12}=6$，$\lambda(v_2)$修改为 1；

同理，把 $T(v_3)$修改为 $P(v_1)+w_{13}=3$，$\lambda(v_3)$修改为 1；把 $T(v_4)$修改为$P(v_1)+w_{14}=1$，$\lambda(v_4)$修改为 1。

转入第三步，在所有的 T 标号中 $T(v_4)=1$ 最小，于是令 $P(v_4)=1$，令 $S_2=S_1\cup\{v_4\}=\{v_1,v_4\}$，$k=4$。

(2) $i=2$

转入第二步，把 $T(v_6)$修改为 $P(v_4)+w_{46}=11$，$\lambda(v_6)$修改为 4。

转入第三步，在所有 T 标号中，$T(v_3)=3$ 最小，于是令 $P(v_3)=3$，令 $S_3=\{v_1,v_4,v_3\}$，$k=3$。

(3) $i=3$

转入第二步，因$(v_3,v_2)\in A$，$v_2\notin S_3$，$T(v_2)>P(v_3)+w_{32}$，把 $T(v_2)$修改为 $P(v_3)+w_{32}=5$，$\lambda(v_2)$修改为 3。

转入第三步，在所有 T 标号中，$T(v_2)=5$ 最小，于是令 $P(v_2)=5, S_4=\{v_1, v_4, v_3, v_2\}, k=2$。

(4) $i=4$

转入第二步，把 $T(v_5)$ 修改为 $P(v_2)+w_{25}=6$，$\lambda(v_5)$ 修改为 2。

转入第三步，在所有 T 标号中，$T(v_5)=6$ 最小，于是令 $P(v_5)=6, S_5=\{v_1, v_4, v_3, v_2, v_5\}, k=5$。

(5) $i=5$

转入第二步，把 $T(v_6), T(v_7), T(v_8)$ 分别修改为 10,9,12，把 $\lambda(v_6), \lambda(v_7), \lambda(v_8)$ 修改为 5。

转入第三步，在所有 T 标号中，$T(v_7)=9$ 最小，于是令

$$P(v_7)=9, \quad S_6=\{v_1, v_4, v_3, v_2, v_5, v_7\}, \quad k=7$$

(6) $i=6$

转入第二步，$(v_7, v_8)\in A, v_8\notin S_6$，但因 $T(v_8)<P(v_7)+w_{73}$，故 $T(v_8)$ 不变。

转入第三步，在所有 T 标号中，$T(v_6)=10$ 最小，令

$$P(v_6)=10, \quad S_7=\{v_1, v_4, v_3, v_2, v_5, v_7, v_6\}, \quad k=6。$$

(7) $i=7$

转入第二步，从 v_6 出发没有弧指向不属于 S_7 的点，故直接转入第 3 步。

转入第三步，在所有 T 标号中，$T(v_8)=12$ 最小，令

$$P(v_8)=12, \quad S_8=\{v_1, v_4, v_3, v_2, v_5, v_7, v_6, v_8\}, \quad k=8。$$

(8) $i=8$

转入第三步，这时仅有的 T 标号点为 v_9，$T(v_9)=+\infty$，故算法终止。

算法终止时

$$P(v_1)=0, \quad P(v_4)=1, \quad P(v_3)=3, \quad P(v_2)=5, \quad P(v_5)=6,$$
$$P(v_7)=9, \quad P(v_6)=10, \quad P(v_8)=12, \quad T(v_9)=+\infty$$
$$\lambda(v_1)=0, \quad \lambda(v_4)=1, \quad \lambda(v_3)=1, \quad \lambda(v_2)=3, \quad \lambda(v_5)=2,$$
$$\lambda(v_7)=5, \quad \lambda(v_6)=5, \quad \lambda(v_8)=5, \quad \lambda(v_9)=M$$

这表示对 $i=1,2,\cdots,8$，$d(v_1, v_i)=P(v_i)$，而从 v_1 到 v_9 不存在路，根据 λ 值可以求出从 v_1 到 v_i 的最短路 $(i=1,2,\cdots,8)$。

例如为了求从 v_1 到 v_8 的最短路，考查 $\lambda(v_8)$，因 $\lambda(v_8)=5$，故最短路包含弧 (v_5, v_8)；再考查 $\lambda(v_5)$，因 $\lambda(v_5)=2$，故最短路包含弧 (v_2, v_5)；类推，$\lambda(v_2)=3$，$\lambda(v_3)=1$，于是最短路包含弧 (v_3, v_2)，及 (v_1, v_3)，这样从 v_1 到 v_8 的最短路是 $(v_1, v_3, v_2, v_5, v_8)$。

上面介绍了求一个赋权有向图中，从一个顶点 v_s 到各个顶点的最短路。对于赋权(无向)图 $G=(V,E)$，因为沿边 $[v_i, v_j]$ 既可以从 v_i 到达 v_j，也可以沿 v_j 到达 v_i，所以边 $[v_i, v_j]$ 可以看做是两条弧 (v_i, v_j) 及 (v_j, v_i)，它们具有相同的权 $w[v_i, v_j]$。这样，在一个

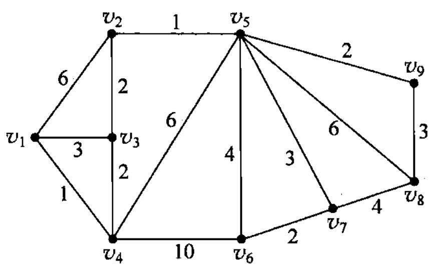


图 11-20

赋权图中，如果所有的 $w_{ij}\geqslant 0$，只要把 Dijkstra 方法中的"第 2 步：考查每个使 $(v_k,v_j)\in A$ 且 $v_j\notin S_i$ 的点 v_j"改为"第 2 步：考查每个使 $[v_k,v_j]\in E$ 且 $v_j\notin S_i$ 的点 v_j"，同样地可以求出从 v_s 到各点的最短路(对于无向图，即为最短链)。

例 11 用 Dijkstra 方法求图 11-20 所示的赋权图中，从 v_1 到 v_8 的最短路。

解 这里只写出计算的最后结果，具体步骤留给读者去完成。

$P(v_1)=0$，$P(v_4)=1$，$P(v_3)=3$，$P(v_2)=5$，$P(v_5)=6$，$P(v_9)=8$，$P(v_7)=9$，$P(v_6)=10$，$P(v_8)=11$。

$\lambda(v_1)=0$，$\lambda(v_4)=1$，$\lambda(v_3)=1$，$\lambda(v_2)=3$，$\lambda(v_5)=2$，$\lambda(v_9)=5$，$\lambda(v_7)=5$，$\lambda(v_6)=5$，$\lambda(v_8)=9$。

这样从 v_1 到 v_8 的最短链为 $(v_1,v_3,v_2,v_5,v_9,v_8)$，总权为 11，从而 $d(v_1,v_8)=11$。

现在来证明 Dijkstra 方法的正确性。只要证明：算法终止时，对于每一个点 $v\in S_i$，$P(v)$ 是从 v_s 到 v 的最短路的权，即 $d(v_s,v)=P(v)$ 即可。

对 i 施行归纳，$i=1$ 时，结论显然正确。设对 $i=n$ 时，结论成立，即对每一个 $v\in S_n$，$d(v_s,v)=P(v)$。现在考查 $i=n+1$，因 $S_{n+1}=S_n\cup\{v_{j_n}\}$，所以只要证明 $d(v_s,v_{j_n})=P(v_{j_n})$。根据算法，$v_{j_n}$ 是这时的具最小 T 标号的点，即

$$T_n(v_{j_n})=\min_{v_j\notin S_n}\{T_n(v_j)\}$$

这里为了清晰起见，用 $T_n(v)$ 表示当 $i=n$ 执行第 3 步时点 v 的 T 标号。假设 H 是 D 中任一条从 v_s 到 v_{j_n} 的路，因为 $v_s\in S_n$，而 $v_{j_n}\notin S_n$，那么从 v_s 出发，沿 H 必存在一条弧，它的始点属于 S_n，而终点不属于 S_n。假设 (v_r,v_l) 是第一条这样的弧，

$$H=(v_s,\cdots,v_r,v_l,\cdots,v_{j_n}),\quad w(H)=w(v_s,\cdots,v_r)+w_{rl}+w(v_l,\cdots,v_{j_n})$$

由归纳假设，$P(v_r)$ 是从 v_s 到 v_r 的最短路的权，于是

$$w(H)\geqslant P(v_r)+w_{rl}+w(v_l,\cdots,v_{j_n})$$

根据方法中 T 标号的修改规则，因 $v_r\in S_n$，$v_l\notin S_n$，故 $P(v_r)+w_{rl}\geqslant T_n(v_l)$。而 $T_n(v_l)\geqslant T_n(v_{j_n})$，故

$$w(H)\geqslant T_n(v_{j_n})+w(v_l,\cdots,v_{j_n})\geqslant T_n(v_{j_n})$$

(因为所有的 $w_{ij}\geqslant 0$，故 $w(v_l,\cdots,v_{j_n})\geqslant 0$)。

这就证明了 $T_n(v_{j_n})$ 是从 v_s 到 v_{j_n} 的最短路的权，由方法，$P(v_{j_n})=T_n(v_{j_n})$，这样就证明了

$$d(v_s,v_{j_n})\doteq P(v_{j_n})$$

Dijkstra 算法只适用于所有 $w_{ij}\geqslant 0$ 的情形，当赋权有向图中存在负权时，则算法失效。例如在如图 11-21 所示的赋

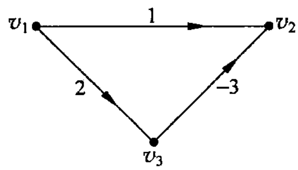


图 11-21

权有向图中，如果用 Dijkstra 方法，可得出从 v_1 到 v_2 的最短路的权是 1，但这显然是不对的，因为从 v_1 到 v_2 的最短路是 (v_1,v_3,v_2)，权是 -1。

现在介绍当赋权有向图 D 中，存在具负权的弧时，求最短路的方法。

为方便起见，不妨设从任一点 v_i 到任一个点 v_j 都有一条弧（如果在 D 中，$(v_i,v_j)\notin A$，则添加弧 (v_i,v_j)。令 $w_{ij}=+\infty$）。

显然，从 v_s 到 v_j 的最短路总是从 v_s 出发，沿着一条路到某个点 v_i，再沿 (v_i,v_j) 到 v_j 的（这里 v_i 可以是 v_s 本身），由本节开始时介绍的一个结论可知，从 v_s 到 v_i 的这条路必定是从 v_s 到 v_i 的最短路，所以 $d(v_s,v_j)$ 必满足如下方程

$$d(v_s,v_j)=\min_i\{d(v_s,v_i)+w_{ij}\}$$

为了求得这个方程的解 $d(v_s,v_1),d(v_s,v_2),\cdots,d(v_s,v_p)$（这里 $p=p(D)$），可用如下递推公式

开始时，令

$$d^{(1)}(v_s,v_j)=w_{sj}\quad(j=1,2,\cdots,p)$$
$$d^{(t)}(v_s,v_j)=\min_i\{d^{(t-1)}(v_s,v_i)+w_{ij}\}$$
$$j=1,2,\cdots,p$$

若进行到某一步，例如第 k 步时，对所有 $j=1,2,\cdots,p$，有

$$d^{(k)}(v_s,v_j)=d^{(k-1)}(v_s,v_j)$$

则 $\{d^{(k)}(v_s,v_j)\}_{j=1,2,\cdots,p}$ 即为 v_s 到各点的最短路的权。

例 12　求图 11-22 所示赋权有向图中从 v_1 到各点的最短路。

解　利用上述递推公式，求解结果如表 11-1 所示（表中未写数字的空格内是 $+\infty$）。

可以看到，当 $t=4$ 时，对所有 $j=1,2,\cdots,8$，有 $d^{(t-1)}(v_1,v_j)=d^{(t)}(v_1,v_j)$，于是表中最后一列 $0,-5,-2,-7,-3,-1,-5,6$ 就分别是从 v_1 到 $v_1,v_2,\cdots,v_8$ 的最短路的权。

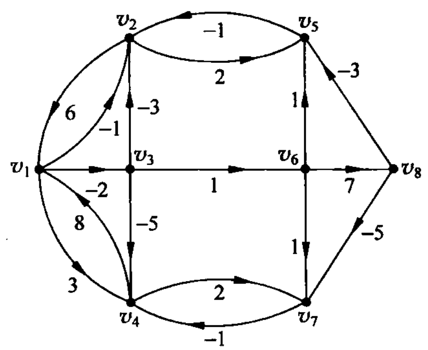


图　11-22

为了进一步求得从 v_s 到各点的最短路，可以类似于 Dijkstra 方法中，给每一个点以 λ 值开始

$$\lambda(v_s)=0,\lambda(v_i)=s\quad(i\neq s)$$

在迭代过程中，如果

$$d^{(t)}(v_s,v_j)=\min\{d^{(t-1)}(v_s,v_i)+w_{ij}\}$$
$$=d^{(t-1)}(v_s,v_{i_0})+w_{i_0 j}$$

则把这时的 $\lambda(v_j)$ 修改为 i_0。迭代终止时，根据各点的 λ 值，可以得到从 v_s 到各点的最短路。

寻求最短路的另一个办法是在求出最短路的权以后，采用“反向追踪”的方法。比如已知 $d(v_s,v_j)$，则寻求一个点 v_k，使 $d(v_s,v_k)+w_{kj}=d(v_s,v_j)$，记录下 (v_k,v_j)，再考查

$d(v_s,v_k)$，寻求一点 v_i，使 $d(v_s,v_i)+w_{ik}=d(v_s,v_k)$，如此等等，直至到达 v_s 为止，于是从 v_s 到 v_j 的最短路是 $(v_s,\cdots,v_i,v_k,v_j)$。

表 11-1

点	w_{ij}								$d^{(t)}(v_1,v_j)$			
	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	$t=1$	$t=2$	$t=3$	$t=4$
v_1	0	−1	−2	3					0	0	0	0
v_2	6	0			2				−1	−5	−5	−5
v_3		−3	0	−5		1			−2	−2	−2	−2
v_4				0			2		3	−7	−7	−7
v_5		−1			0					1	−3	−3
v_6					1	0	1	7		−1	−1	−1
v_7				−1			0			5	−5	−5
v_8					−3		−5	0			6	6

如例 3，由表 11-1 已知，$d(v_1,v_8)=6$，

因 $d(v_1,v_6)+w_{68}=(-1)+7=d(v_1,v_8)$，故记下 (v_6,v_8)。

因 $d(v_1,v_3)+w_{36}=d(v_1,v_6)$，故记下 (v_3,v_6)。

因 $d(v_1,v_7)+w_{73}=d(v_1,v_3)$，从而从 v_1 到 v_8 的最短路是 (v_1,v_3,v_6,v_8)。

定义 5 设 D 是赋权有向图，C 是 D 中的一个回路，如果 C 的权 $w(C)$ 小于零，则称 C 是 D 中的一个**负回路**。

不难证明：

(1) 如果 D 是不含负回路的赋权有向图，那么，从 v_s 到任一个点的最短路必可取为初等路，从而最多包含 $p-2$ 个中间点；

(2) 上述递推公式中的 $d^{(t)}(v_s,v_j)$ 是在至多包含 $t-1$ 个中间点的限制条件下，从 v_s 到 v_j 的最短路的权。

由(1)、(2)可知：当 D 中不含负回路时，上述算法最多经过 $p-1$ 次迭代必定收敛，即对所有的 $j=1,2,\cdots,p$，均有 $d^{(k)}(v_s,v_j)=d^{(k-1)}(v_s,v_j)$，从而求出从 v_s 到各个顶点的最短路的权。

如果经过 $p-1$ 次迭代，存在某个 j，使 $d^{(p)}(v_s,v_j)\neq d^{(p-1)}(v_s,v_j)$，则说明 D 中含有负回路。显然，这时从 v_s 到 v_j 的路的权是没有下界的。

为了加快收敛速度，可以利用如下的递推公式。

$$d^{(1)}(v_s,v_j)=w_{sj}\quad (j=1,2,\cdots,p)$$

$$d^{(t)}(v_s,v_j)=\min\{\min_{i<j}\{d^{(t)}(v_s,v_j)+w_{ij}\},\min_{i\geqslant j}\{d^{(t-1)}(v_s,v_i)+w_{ij}\}\}$$

$$(j=1,2,\cdots,p),(t=2,3,\cdots)$$

J. Y. Yen 提出一个改进的递推算法：

$$d^{(1)}(v_s,v_j)=w_{sj}\quad j=1,2,\cdots,p$$

对 $t=2,4,6,\cdots$，按 $j=1,2,\cdots,p$ 的顺序计算：

$$d^{(t)}(v_s,v_j)=\min\{d^{(t-1)}(v_s,v_j),\min_{i<j}\{d^{(t)}(v_s,v_i)+w_{ij}\}\}$$

对 $t=3,5,7,\cdots$，按 $j=p,p-1,\cdots,1$ 的顺序计算：

$$d^{(t)}(v_s,v_j)=\min\{d^{(t-1)}(v_s,v_j),\min_{i>j}\{d^{(t)}(v_s,v_i)+w_{ij}\}\}$$

同样地，当对所有的 $j=1,2,\cdots,p$

$$d^{(k)}(v_s,v_j)=d^{(k-1)}(v_s,v_j)$$

时，算法终止。

11.3.3 应用举例

例 13 设备更新问题。某企业使用一台设备，在每年年初，企业领导部门就要决定是购置新的，还是继续使用旧的。若购置新设备，就要支付一定的购置费用；若继续使用旧设备，则需支付一定的维修费用。现在的问题是如何制定一个几年之内的设备更新计划，使得总的支付费用最少。我们用一个五年之内要更新某种设备的计划为例，若已知该种设备在各年年初的价格如表 11-2 所示：

表 11-2

第 1 年	第 2 年	第 3 年	第 4 年	第 5 年
11	11	12	12	13

还已知使用不同时间(年)的设备所需要的维修费用如表 11-3 所示：

表 11-3

使用年数	0～1	1～2	2～3	3～4	4～5
维修费用	5	6	8	11	18

可供选择的设备更新方案显然是很多的。例如，每年都购置一台新设备，则其购置费用为 11＋11＋12＋12＋13＝59，而每年支付的维修费用为 5，五年合计为 25。于是五年总的支付费用为 59＋25＝84。

又如决定在第一、三、五年各购进一台，这个方案的设备购置费为 11＋12＋13＝36，维修费为 5＋6＋5＋6＋5＝27。五年总的支付费用为 63。

如何制定使得总的支付费用最少的设备更新计划呢？可以把这个问题化为最短路问题，见图 11-23。

用点 v_i 代表“第 i 年年初购进一台新设备”这种状态（加设一点 v_6，可以理解为第 5 年年底）。从 v_i 到 $v_{i+1},\cdots,v_6$ 各画一条弧。弧(v_i,v_j)表示在第 i 年年初购进的设备一直使用到第 j 年年初（即第 $j-1$ 年年底）。

每条弧的权可按已知资料计算出来。例如，(v_1,v_4)是第 1 年年初购进一台新设备

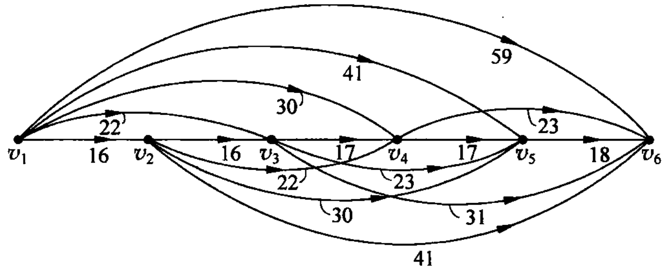


图 11-23

(支付购置费 11),一直使用到第 3 年年底(支付维修费 5+6+8=19),故(v_1,v_4)上的权为 30。

这样一来,制定一个最优的设备更新计划问题就等价于寻求从 v_1 到 v_6 的最短路的问题。

按求解最短路的计算方法,$\{v_1,v_3,v_6\}$及$\{v_1,v_4,v_6\}$均为最短路,即有两个最优方案。一个方案是在第 1 年、第 3 年各购置一台新设备;另一个方案是在第 1 年、第 4 年各购置一台新设备。五年总的支付费用均为 53。

11.4 网络最大流问题

许多系统包含了流量问题。例如,公路系统中有车辆流,控制系统中有信息流,供水系统中有水流,金融系统中有现金流等。

图 11-24 是联结某产品产地 v_1 和销地 v_6 的交通网,每一弧(v_i,v_j)代表从 v_i 到 v_j 的运输线,产品经这条弧由 v_i 输送到 v_j,弧旁的数字表示这条运输线的最大通过能力。产品经过交通网从 v_1 输送到 v_6。现在要求制定一个运输方案使从 v_1 运到 v_6 的产品数量最多。

图 11-25 给出了一个运输方案,每条弧旁的数字表示在这个方案中,每条运输线上的运输数量。这个方案使 8 个单位的产品从 v_1 运到 v_6,在这个交通网上输送量是否还可以增多,或者说这个运输网络中,从 v_1 到 v_6 的最大输送量是多少呢?本节就是要研究类似这样的问题。

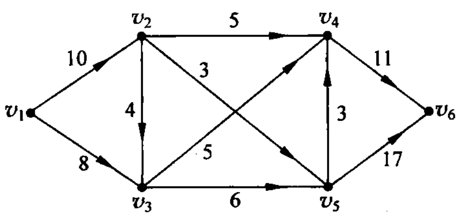


图 11-24

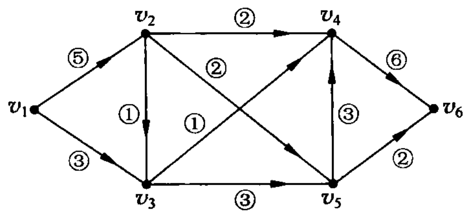


图 11-25

11.4.1 基本概念与基本定理

1. 网络与流

定义 6 给一个有向图 $D=(V,A)$，在 V 中指定了一点称为**发点**(记为 v_s)，而另一点称为**收点**(记为 v_t)，其余的点叫中间点。对于每一个弧 $(v_i,v_j)\in A$，对应有一个 $c(v_i,v_j)\geqslant 0$(或简写为 c_{ij})，称为弧的**容量**。通常我们就把这样的 D 叫做一个**网络**。记作

$$D=(V,A,C)$$

所谓网络上的流，是指定义在弧集合 A 上的一个函数 $f=\{f(v_i,v_j)\}$，并称 $f(v_i,v_j)$ 为弧 (v_i,v_j) 上的**流量**(有时也简记作 f_{ij})。

例如，图 11-24 就是一个网络，指定 v_1 是发点，v_6 是收点，其他的点是中间点。弧旁的数字为 c_{ij}。

图 11-25 所示的运输方案，就可看做是这个网络上的一个流，每个弧上的运输量就是该弧上的流量，即 $f_{12}=5$， $f_{24}=2$， $f_{13}=3$， $f_{34}=1$ 等。

2. 可行流与最大流

在运输网络的实际问题中可以看出，对于流有两个明显的要求：一是每个弧上的流量不能超过该弧的最大通过能力(即弧的容量)；二是中间点的流量为零。因为对于每个点，运出这点的产品总量与运进这点的产品总量之差，是这点的净输出量，简称为是这一点的流量；由于中间点只起转运作用，所以中间点的流量必为零。易见发点的净流出量和收点的净流入量必相等，也是这个方案的总输送量。因此有：

定义 7 满足下述条件的流 f 称为**可行流**：

(1) 容量限制条件：对每一弧 $(v_i,v_j)\in A$

$$0\leqslant f_{ij}\leqslant c_{ij}$$

(2) 平衡条件：

对于中间点：流出量等于流入量，即对每个 $i(i\neq s,t)$ 有

$$\sum_{(v_i,v_j)\in A} f_{ij}-\sum_{(v_j,v_i)\in A} f_{ji}=0$$

对于发点 v_s，记

$$\sum_{(v_s,v_j)\in A} f_{sj}-\sum_{(v_j,v_s)\in A} f_{js}=v(f)$$

对于收点 v_t，记

$$\sum_{(v_t,v_j)\in A} f_{tj}-\sum_{(v_j,v_t)\in A} f_{jt}=-v(f)$$

式中，$v(f)$ 称为这个可行流的流量，即发点的净输出量(或收点的净输入量)。

可行流总是存在的。比如令所有弧的流量 $f_{ij}=0$，就得到一个可行流(称为**零流**)。其流量 $v(f)=0$。

最大流问题就是求一个流 $\{f_{ij}\}$ 使其流量 $v(f)$ 达到最大，并且满足

$$0\leqslant f_{ij}\leqslant c_{ij} \qquad (v_i,v_j)\in A \qquad ①$$

$$\sum f_{ij}-\sum f_{ji}=\begin{cases}v(f) & (i=s)\\ 0 & (i\neq s,t)\\ -v(f) & (i=t)\end{cases} \qquad ②$$

最大流问题是一个特殊的线性规划问题。即求一组$\{f_{ij}\}$，在满足条件①和②下使$v(f)$达到极大。将会看到利用图的特点，解决这个问题的方法较之线性规划的一般方法要方便、直观得多。

3. 增广链

若给一个可行流$f=\{f_{ij}\}$，我们把网络中使$f_{ij}=c_{ij}$的弧称为**饱和弧**，使$f_{ij}<c_{ij}$的弧称为**非饱和弧**。使$f_{ij}=0$的弧称为**零流弧**，使$f_{ij}>0$的弧称为**非零流弧**。

在图11-25中，(v_5,v_4)是饱和弧，其他的弧为非饱和弧。所有弧都是非零流弧。

若μ是网络中联结发点v_s和收点v_t的一条链，我们定义链的方向是从v_s到v_t，则链上的弧被分为两类：一类是弧的方向与链的方向一致，叫做**前向弧**。前向弧的全体记为μ^+。另一类弧与链的方向相反，称为**后向弧**。后向弧的全体记为μ^-。

图11-24中，在链$\mu=(v_1,v_2,v_3,v_4,v_5,v_6)$上

$$\mu^+=\{(v_1,v_2),(v_2,v_3),(v_3,v_4),(v_5,v_6)\}$$

$$\mu^-=\{(v_5,v_4)\}$$

定义8 设f是一个可行流，μ是从v_s到v_t的一条链，若μ满足下列条件，称之为(关于可行流f的)**增广链**。

在弧$(v_i,v_j)\in\mu^+$上，$0\leqslant f_{ij}<c_{ij}$，即μ^+中每一弧是非饱和弧。

在弧$(v_i,v_j)\in\mu^-$上，$0<f_{ij}\leqslant c_{ij}$，即μ^-中每一弧是非零流弧。

图11-25中，链$\mu=(v_1,v_2,v_3,v_4,v_5,v_6)$是一条增广链。因为$\mu^+$和$\mu^-$中的弧满足增广链的条件。比如：

$$(v_1,v_2)\in\mu^+,\quad f_{12}=5<c_{12}=10$$

$$(v_5,v_4)\in\mu^-,\quad f_{54}=3>0$$

4. 截集与截量

设$S,T\subset V,S\cap T=\varnothing$，我们把始点在$S$中，终点在$T$中的所有弧构成的集合，记为$(S,T)$。

定义9 给网络$D=(V,A,C)$，若点集V被剖分为两个非空集合V_1和$\bar{V}_1$，使$v_s\in V_1,v_t\in\bar{V}_1$，则把弧集$(V_1,\bar{V}_1)$称为是(分离$v_s$和$v_t$的)**截集**。

显然，若把某一截集的弧从网络中丢去，则从v_s到v_t便不存在路。所以，直观上说，截集是从v_s到v_t的必经之道。

定义10 给一截集$(V_1,\bar{V}_1)$，把截集$(V_1,\bar{V}_1)$中所有弧的容量之和称为这个截集的容量(简称为**截量**)，记为$c(V_1,\bar{V}_1)$，即

$$c(V_1,\bar{V}_1)=\sum_{(v_i,v_j)\in(V_1,\bar{V}_1)}c_{ij}$$

不难证明，任何一个可行流的流量 $v(f)$ 都不会超过任一截集的容量。即

$$v(f)\leqslant c(V_1,\bar{V}_1)$$

显然，若对于一个可行流 f^*，网络中有一个截集 $(V_1^*,\bar{V}_1^*)$，使 $v(f^*)=c(V_1^*,\bar{V}_1^*)$，则 f^* 必是最大流，而 $(V_1^*,\bar{V}_1^*)$ 必定是 D 的所有截集中，容量最小的一个，即最小截集。

定理 8　可行流 f^* 是最大流，当且仅当不存在关于 f^* 的增广链。

证明　若 f^* 是最大流，设 D 中存在关于 f^* 的增广链 μ，令

$$\theta=\min\{\min_{\mu^+}(c_{ij}-f_{ij}^*),\min_{\mu^-}f_{ij}^*\}$$

由增广链的定义，可知 $\theta>0$，令

$$f_{ij}^{**}=\begin{cases}f_{ij}^*+\theta & (v_i,v_j)\in\mu^+\\ f_{ij}^*-\theta & (v_i,v_j)\in\mu^-\\ f_{ij}^* & (v_i,v_j)\notin\mu\end{cases}$$

不难验证 $\{f_{ij}^{**}\}$ 是一个可行流，且 $v(f^{**})=v(f^*)+\theta>v(f^*)$。这与 f^* 是最大流的假设矛盾。

现在设 D 中不存在关于 f^* 的增广链，证明 f^* 是最大流。我们利用下面的方法来定义 V_1^*：

令 $v_s\in V_1^*$

若 $v_i\in V_1^*$，且 $f_{ij}^*<c_{ij}$，则令 $v_j\in V_1^*$

若 $v_i\in V_1^*$，且 $f_{ji}^*>0$，　则令 $v_j\in V_1^*$

因为不存在关于 f^* 的增广链，故 $v_t\notin V_1^*$。

记 $\bar{V}_1^*=V\backslash V_1^*$，于是得到一个截集 $(V_1^*,\bar{V}_1^*)$。显然必有

$$f_{ij}^*=\begin{cases}c_{ij} & (v_i,v_j)\in(V_1^*,\bar{V}_1^*)\\ 0 & (v_i,v_j)\in(\bar{V}_1^*,V_1^*)\end{cases}$$

所以 $v(f^*)=c(V_1^*,\bar{V}_1^*)$。于是 f^* 必是最大流。定理得证。

由上述证明中可见，若 f^* 是最大流，则网络中必存在一个截集 $(V_1^*,\bar{V}_1^*)$，使

$$v(f^*)=C(V_1^*,\bar{V}_1^*)$$

于是有如下重要的结论：

最大流量最小截量定理：任一个网络 D 中，从 v_s 到 v_t 的最大流的流量等于分离 v_s，v_t 的最小截集的容量。

定理 8 为我们提供了寻求网络中最大流的一个方法。若给了一个可行流 f，只要判断 D 中有无关于 f 的增广链。如果有增广链，则可以按定理 8 前半部证明中的办法，改进 f，得到一个流量增大的新的可行流。如果没有增广链，则得到最大流。而利用定理 8 后半部证明中定义 V_1^* 的办法，可以根据 v_t 是否属于 V_1^* 来判断 D 中有无关于 f 的增广链。

实际计算时，用给顶点标号的方法来定义 V_1^*。在标号过程中，有标号的顶点表示是 V_1^* 中的点，没有标号的点表示不是 V_1^* 中的点。一旦 v_t 有了标号，就表明找到一条增广链；如果标号过程进行不下去，而 v_t 尚未标号，则说明不存在增广链，于是得到最大流。

而且同时也得到一个最小截集。

11.4.2 寻求最大流的标号法

从一个可行流出发(若网络中没有给定 f,则可以设 f 是零流),经过标号过程与调整过程。

1. 标号过程

在这个过程中,网络中的点或者是标号点(又分为已检查和未检查两种),或者是未标号点。每个标号点的标号包含两部分:第一个标号表明它的标号是从哪一点得到的,以便找出增广链;第二个标号是为确定增广链的调整量 θ 用的。

标号过程开始,总先给 v_s 标上 $(0,+\infty)$,这时 v_s 是标号而未检查的点,其余都是未标号点。一般地,取一个标号而未检查的点 v_i,对一切未标号点 v_j:

(1) 若在弧 (v_i,v_j) 上,$f_{ij}<c_{ij}$,则给 v_j 标号 $(v_i,l(v_j))$。这里 $l(v_j)=\min[l(v_i),c_{ij}-f_{ij}]$。这时点 v_j 成为标号而未检查的点。

(2) 若在弧 (v_j,v_i) 上,$f_{ji}>0$,则给 v_j 标号 $(-v_i,l(v_j))$,这里 $l(v_j)=\min[l(v_i),f_{ji}]$。这时点 v_j 成为标号而未检查的点。

于是 v_i 成为标号而已检查过的点。重复上述步骤,一旦 v_t 被标上号,表明得到一条从 v_s 到 v_t 的增广链 μ,转入调整过程。

若所有标号都是已检查过的,而标号过程进行不下去时,则算法结束,这时的可行流就是最大流。

2. 调整过程

首先按 v_t 及其他点的第一个标号,利用"反向追踪"的办法,找出增广链 μ。例如设 v_t 的第一个标号为 v_k(或 $-v_k$),则弧 (v_k,v_t)(或相应地 (v_t,v_k))是 μ 上的弧。接下来检查 v_k 的第一个标号,若为 v_i(或 $-v_i$),则找出 (v_i,v_k)(或相应地 (v_k,v_i))。再检查 v_i 的第一个标号,依此下去,直到 v_s 为止。这时被找出的弧就构成了增广链 μ。令调整量 θ 是 $l(v_t)$,即 v_t 的第二个标号。令

$$f'_{ij}=\begin{cases} f_{ij}+\theta & (v_i,v_j)\in\mu^+ \\ f_{ij}-\theta & (v_i,v_j)\in\mu^- \\ f_{ij} & (v_i,v_j)\notin\mu \end{cases}$$

去掉所有的标号,对新的可行流 $f'=\{f'_{ij}\}$,重新进入标号过程。

例 14 用标号法求图 11-26 所示网络的最大流。弧旁的数是 (c_{ij},f_{ij})。

解 (1) 标号过程

① 首先给 v_s 标上 $(0,+\infty)$

② 检查 v_s,在弧 (v_s,v_2) 上,$f_{s2}=c_{s2}=3$,不满足标号条件。弧 (v_s,v_1) 上,$f_{s1}=1$,$c_{s1}=5$,$f_{s1}<c_{s1}$,则 v_1 的标号为 $(v_s,l(v_1))$,其中

$$l(v_1)=\min[l(v_s),(c_{s1}-f_{s1})]=\min[+\infty,5-1]=4$$

③ 检查 v_1,在弧 (v_1,v_3) 上,$f_{13}=2$,$c_{13}=2$,不满足标号条件。

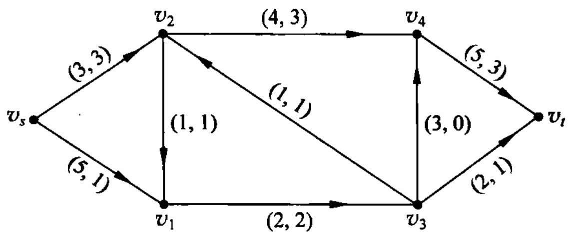


图　11-26

在弧(v_2,v_1)上，$f_{21}=1>0$，则给 v_2 记下标号为$(-v_1,l(v_2))$，这里

$$l(v_2)=\min[l(v_1),f_{21}]=\min[4,1]=1$$

④ 检查 v_2，在弧(v_2,v_4)上，$f_{24}=3$，$c_{24}=4$，$f_{24}<c_{24}$，则给 v_4 标号$(v_2,l(v_4))$，这里

$$l(v_4)=\min[l(v_2),(c_{24}-f_{24})]=\min[1,1]=1$$

在弧(v_3,v_2)上，$f_{32}=1>0$，给 v_3 标号：$(-v_2,l(v_3))$，这里

$$l(v_3)=\min[l(v_2),f_{32}]=\min[1,1]=1$$

⑤ 在 v_3，v_4 中任选一个进行检查。例如，

在弧(v_3,v_t)上，$f_{3t}=1$，$c_{3t}=2$，$f_{3t}<c_{3t}$，给 v_t 标号为$(v_3,l(v_t))$，这里

$$l(v_t)=\min[l(v_3),(c_{3t}-f_{3t})]=\min[1,1]=1$$

因 v_t 有了标号，故转入调整过程。

(2) 调整过程

按点的第一个标号找到一条增广链，如图 11-27 中双箭头线表示。

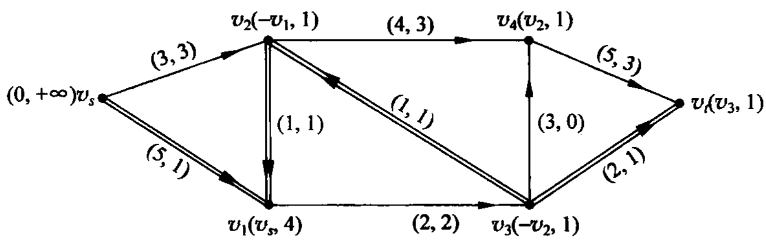


图　11-27

易见

$$\mu^+=\{(v_s,v_1),(v_3,v_t)\}$$

$$\mu^-=\{(v_2,v_1),(v_3,v_2)\}$$

按 $\theta=1$ 在 μ 上调整 f。

μ^+ 上：　$f_{s1}+\theta=1+1=2$

$f_{3t}+\theta=1+1=2$

μ^- 上：　$f_{21}-\theta=1-1=0$

$f_{32}-\theta=1-1=0$

其余的 f_{ij} 不变。

调整后得如图 11-28 所示的可行流，对这个可行流进入标号过程，寻找增广链。

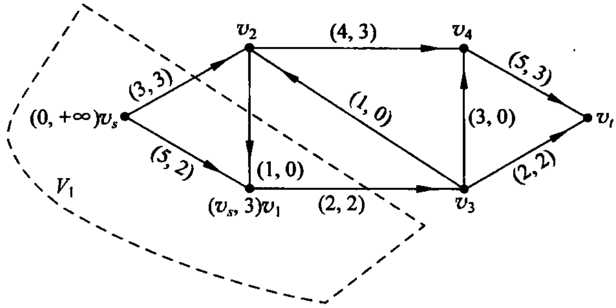


图 11-28

开始给 v_s 标以$(0,+\infty)$，于是检查 v_s，给 v_1 标以$(v_s,3)$，检查 v_1，弧(v_1,v_3)上，$f_{13}=c_{13}$，弧(v_2,v_1)上，$f_{21}=0$，均不符号条件，标号过程无法继续下去，算法结束。

这时的可行流(图 11-28)即为所求最大流。最大流量为

$$v(f)=f_{s1}+f_{s2}=f_{4t}+f_{3t}=5$$

与此同时可找到最小截集$(V_1,\bar{V}_1)$，其中 V_1 为标号点集合，$\bar{V}_1$ 为未标号点集合。弧集合$(V_1,\bar{V}_1)$即为最小截集。

上例中，$V_1=\{v_s,v_1\}$，$\bar{V}_1=\{v_2,v_3,v_4,v_t\}$，于是$(V_1,\bar{V}_1)=\{(v_s,v_2),(v_1,v_3)\}$是最小截集，它的容量也是 5。

由上述可见，用标号法找增广链以求最大流的结果，同时得到一个最小截集。最小截集容量的大小影响总的输送量的提高。因此，为提高总的输送量，必须首先考虑改善最小截集中各弧的输送状况，提高它们的通过能力。另一方面，一旦最小截集中弧的通过能力被降低，就会使总的输送量减少。

11.5 最小费用最大流问题

上一节讨论了寻求网络中的最大流问题。在实际生活中，涉及“流”的问题时，人们考虑的还不只是流量，而且还有“费用”的因素，本节介绍的最小费用最大流问题就是这类问题之一。

给网络 $D=(V,A,C)$，每一弧$(v_i,v_j)\in A$ 上，除了已给容量 c_{ij} 外，还给了一个单位流量的费用 $b(v_i,v_j)\geqslant 0$(简记为 b_{ij})。所谓**最小费用最大流问题**就是要求一个最大流 f，使流的总输送费用

$$b(f)=\sum_{(v_i,v_j)\in A} b_{ij}f_{ij}$$

取极小值。

下面介绍解决这个问题的一种方法。

从上节可知，寻求最大流的方法是从某个可行流出发，找到关于这个流的一条增广链 μ。沿着 μ 调整 f，对新的可行流试图寻求关于它的增广链，如此反复直至最大流。现在要寻求最小费用的最大流，首先考查一下，当沿着一条关于可行流 f 的增广链 μ，以 $\theta=1$ 调整 f，得到新的可行流 f' 时（显然 $v(f')=v(f)+1$），$b(f')$ 比 $b(f)$ 增加多少？不难看出

$$b(f')-b(f)=\left[\sum_{\mu^+} b_{ij}(f'_{ij}-f_{ij})-\sum_{\mu^-} b_{ij}(f'_{ij}-f_{ij})\right]$$
$$=\sum_{\mu^+} b_{ij}-\sum_{\mu^-} b_{ij}$$

我们把 $\sum_{\mu^+} b_{ij}-\sum_{\mu^-} b_{ij}$ 称为这条增广链 μ 的“费用”。

可以证明，若 f 是流量为 $v(f)$ 的所有可行流中费用最小者，而 μ 是关于 f 的所有增广链中费用最小的增广链，那么沿 μ 去调整 f，得到的可行流 f'，就是流量为 $v(f')$ 的所有可行流中的最小费用流。这样，当 f' 是最大流时，它也就是所要求的最小费用最大流了。

注意到，由于 $b_{ij}\geqslant 0$，所以 $f=0$ 必是流量为 0 的最小费用流。这样，总可以从 $f=0$ 开始。一般地，设已知 f 是流量 $v(f)$ 的最小费用流，余下的问题就是如何去寻求关于 f 的最小费用增广链。为此，可构造一个赋权有向图 $W(f)$，它的顶点是原网络 D 的顶点，而把 D 中的每一条弧 (v_i,v_j) 变成两个相反方向的弧 (v_i,v_j) 和 (v_j,v_i)。定义 $W(f)$ 中弧的权 w_{ij} 为

$$w_{ij}=\begin{cases} b_{ij} & 若\ f_{ij}<c_{ij} \\ +\infty & 若\ f_{ij}=c_{ij} \end{cases}$$

$$w_{ji}=\begin{cases} -b_{ij} & 若\ f_{ij}>0 \\ +\infty & 若\ f_{ij}=0 \end{cases}$$

（长度为 $+\infty$ 的弧可以从 $W(f)$ 中略去）

于是在网络 D 中寻求关于 f 的最小费用增广链就等价于在赋权有向图 $W(f)$ 中，寻求从 v_s 到 v_t 的最短路。因此有如下算法。

开始取 $f^{(0)}=0$，一般情况下若在第 $k-1$ 步得到最小费用流 $f^{(k-1)}$，则构造赋权有向图 $W(f^{(k-1)})$，在 $W(f^{(k-1)})$ 中，寻求从 v_s 到 v_t 的最短路。若不存在最短路（即最短路权是 $+\infty$），则 $f^{(k-1)}$ 就是最小费用最大流；若存在最短路，则在原网络 D 中得到相应的增广链 μ，在增广链 μ 上对 $f^{(k-1)}$ 进行调整。调整量为

$$\theta=\min\left[\min_{\mu^+}(c_{ij}-f_{ij}^{(k-1)}),\min_{\mu^-}(f_{ij}^{(k-1)})\right]$$

令

$$f_{ij}^{(k)}=\begin{cases} f_{ij}^{(k-1)}+\theta & (v_i,v_j)\in\mu^+ \\ f_{ij}^{(k-1)}-\theta & (v_i,v_j)\in\mu^- \\ f_{ij}^{(k-1)} & (v_i,v_j)\notin\mu \end{cases}$$

得到新的可行流 $f^{(k)}$，再对 $f^{(k)}$ 重复上述步骤。

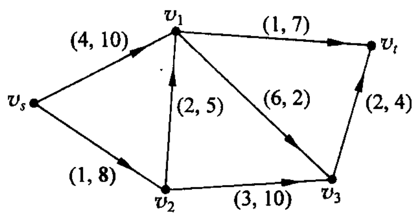


图 11-29

例 15 以图 11-29 为例，求最小费用最大流。弧旁数字为(b_{ij}, c_{ij})。

(1) 取 $f^{(0)}=0$ 为初始可行流。

(2) 构造赋权有向图 $W(f^{(0)})$，并求出从 v_s 到 v_t 的最短路(v_s, v_2, v_1, v_t)，如图 11-30(a)(双箭头即为最短路)。

(3) 在原网络 D 中，与这条最短路相应的增广链为 $\mu=(v_s, v_2, v_1, v_t)$。

(4) 在 μ 上进行调整，$\theta=5$，得 $f^{(1)}$(图 11-30(b))。按照上述算法依次得 $f^{(1)}, f^{(2)}, f^{(3)}, f^{(4)}$，流量依次为 5,7,10,11；构造相应的赋权有向图为 $W(f^{(1)}), W(f^{(2)}), W(f^{(3)}), W(f^{(4)})$，见图 11-30。

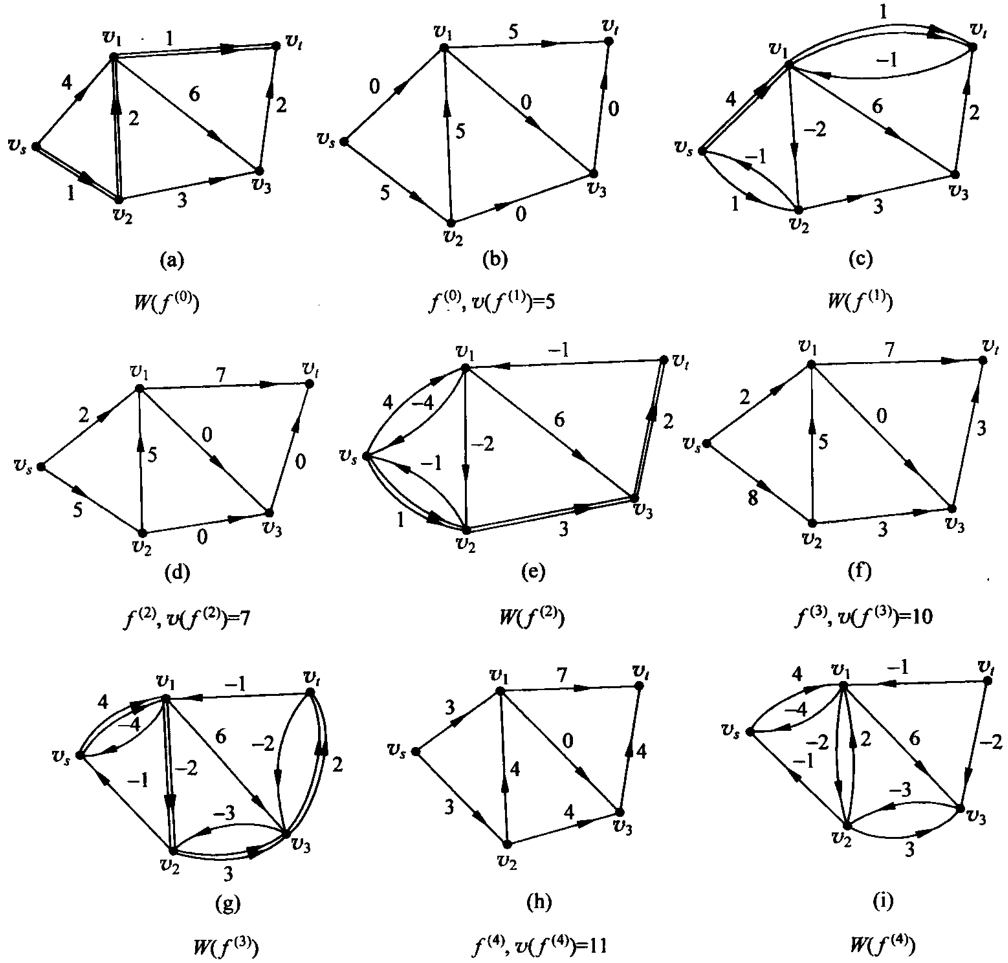


图 11-30

注意到 $W(f^{(4)})$ 中已不存在从 v_s 到 v_t 的最短路，所以 $f^{(4)}$ 为最小费用最大流。

11.6　中国邮递员问题

在本章开始提到的邮递员问题，若把它抽象为图的语言，就是给定一个连通图，在每边 e_i 上赋予一个非负的权 $w(e_i)$，要求一个圈（未必是简单的），过每边至少一次，并使圈的总权最小。这个问题是我国学者管梅谷在 1962 年首先提出的，因此在国际上通称为**中国邮递员问题**。

11.6.1　一笔画问题

给定一个连通多重图 G，若存在一条链，过每边一次，且仅一次，则称这条链为**欧拉链**。若存在一个简单圈，过每边一次，且仅一次，称这个圈为**欧拉圈**。一个图若有欧拉圈，则称为**欧拉图**。显然，一个图若能一笔画出，这个图必是欧拉图（出发点与终止点重合）或含有欧拉链（出发点与终止点不同）。

定理 9　连通多重图 G 有欧拉圈，当且仅当 G 中无奇点。

证明　必要性是显然的，只证明充分性。不妨设 G 至少有三个点，对边数 $q(G)$ 进行数学归纳，因 G 是连通图，不含奇点，故 $q(G)\geqslant 3$。首先 $q(G)=3$ 时，G 显然是欧拉图。考查 $q(G)=n+1$ 的情况，因 G 是不含奇点的连通图，并且 $p(G)\geqslant 3$，故存在三个点 u,v,w，使 $[u,v]$，$[w,v]\in E$。从 G 中丢去边 $[u,v]$，$[w,v]$，增加新边 $[u,w]$，得到新的多重图 G'。G' 有 $q(G)-1$ 条边，并且仍不含奇点，G' 至多有两个分图。若 G' 是连通的，那么根据归纳假设，G' 有欧拉圈 C'。把 C' 中的 $[w,u]$ 这一条边换成 $[w,v]$，$[v,u]$，即得 G 中的欧拉圈。现设 G' 有两个分图 G_1，G_2。设 v 在 G_1 中。根据归纳假设，G_1，G_2 分别有欧拉圈 C_1，C_2，则把 C_2 中的 $[u,w]$ 这条边换成 $[u,v]$，C_1 及 $[v,w]$，即得 G 的欧拉圈。

推论　连通多重图 G 有欧拉链，当且仅当 G 恰有两个奇点。

证明　必要性是显然的。现设连通多重图 G 恰有两个奇点 u，v。在 G 中增加一个新边 $[u,v]$（如果在 G 中，u，v 之间就有边，那么这个新边是原有边上的重复边），得连通多重图 G'，易见 G' 中无奇点。由定理 3，G' 有欧拉圈 C'，从 C' 中丢去增加的那个新边 $[u,v]$，即得 G 中的一条联结 u，v 的欧拉链。

上述定理和推论为我们提供了识别一个图能否是一笔画的简单办法。如前面提到的七桥问题，因为图 11-1(b)中有 4 个奇点。所以不能一笔画出。也就是说，七桥问题的回答是否定的，如图 11-31。它有两个奇点 v_2 和 v_5，因此可以从点 v_2 开始，用一笔画到点 v_5 终止。

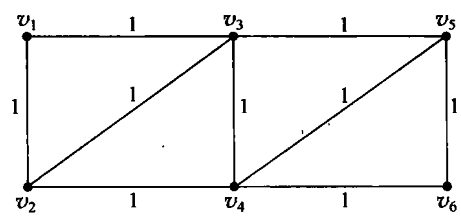


图　11-31

现在的问题是：如果我们已经知道图 G 是

可以一笔画的，怎么样把它一笔画出来呢？也就是说，怎么找出它的欧拉圈(这时 G 无奇点)或欧拉链(这时 G 恰有两个奇点)呢？下面简单地介绍由 Fleury 提供的方法。

为此，首先介绍割边的概念。设 e 是连通图 G 的一个边，如果从 G 中丢去 e，图就不连通了，则称 e 是图 G 的**割边**。例如，图 11-10(b)中，$[v_1,v_2]$是割边；树中的每一个边都是割边。

设 $G=(V,E)$是无奇点的连通图，以

$$\mu_k=(v_{i_0},e_{i_1},v_{i_1},e_{i_2},v_{i_2},\cdots,v_{i_{k-1}},e_{i_k},v_{i_k})$$

记在第 k 步得到的简单链。记 $E_k=\{e_{i_1},e_{i_2},\cdots,e_{i_k}\}$，$\bar{E}_k=E\backslash E_k$，以及 $G_k=(V,\bar{E}_k)$(开始 $k=0$时，令 $\mu_0=(v_{i_0})$，这里 v_{i_0} 是图 G 的任意一点，$E_0=\varnothing$；$G_0=G$)。进行第$(k+1)$步：在 G_k 中选 v_{i_k} 的一条关联边 $e_{i_{k+1}}=[v_{i_k},v_{i_{k+1}}]$，使 $e_{i_{k+1}}$ 不是 G_k 的割边(除非 v_{i_k} 是 G_k 的悬挂点，这时 v_{i_k} 在 G_k 中的悬挂边选为 $e_{i_{k+1}}$)。令

$$\mu_{k+1}=(v_{i_0},e_{i_1},v_{i_1},e_{i_2},v_{i_2},\cdots,v_{i_{k-1}},e_{i_k},v_{i_k},e_{i_{k+1}},v_{i_{k+1}})$$

重复这个过程，直到选不到所要求的边为止。可以证明：这时的简单链必定终止于 v_{i0}，并且就是我们要求的图 G 的欧拉圈。

如果 $G=(V,E)$是恰有两个奇点的连通图。只需要取 v_{i0} 是图 G 的一个奇点就可以了。最终得到的简单链就是图中联结两个奇点的欧拉链。

11.6.2 奇偶点图上作业法

根据上面的讨论，如果在某邮递员所负责的范围内，街道图中没有奇点，那么他就可以从邮局出发，走过每条街道一次，且仅一次，最后回到邮局，这样他所走的路程也就是最短的路程。对于有奇点的街道图，就必须在某些街道上重复走一次或多次。

例如图 11-31 的街道图中，若 v_1 是邮局，邮递员可以按如下的路线投递信件：

$v_1\to v_2\to v_4\to v_3\to v_2\to v_4\to v_6\to v_5\to v_4\to v_6\to v_5\to v_3\to v_1$，总权为 12。

也可按另一条路线走：

$v_1\to v_2\to v_3\to v_2\to v_4\to v_5\to v_6\to v_4\to v_3\to v_5\to v_3\to v_1$，总权为 11。

可见，按第一条路线走，在边$[v_2,v_4]$，$[v_4,v_6]$，$[v_6,v_5]$上各重复走了一次。而按第二条路线走，在边$[v_3,v_2]$，$[v_3,v_5]$上各重复走了一次。

如果在某条路线中，边$[v_i,v_j]$上重复走了几次，我们在图中 v_i，v_j 之间增加几条边，令每条边的权和原来的权相等，并把新增加的边，称为重复边。于是这条路线就是相应的新图中的欧拉圈。例如在图 11-31 中，上面提到的两条投递路线分别是图 11-32(a)和(b)中的欧拉圈。

显然，两条邮递路线的总权的差必等于相应的重复边总权的差。因而，中国邮递员问题可以叙述为在一个有奇点的图中，要求增加一些重复边，使新图不含奇点，并且重复边的总权为最小。

我们把使新图不含奇点而增加的重复边，简称为可行(重复边)方案，使总权最小的可

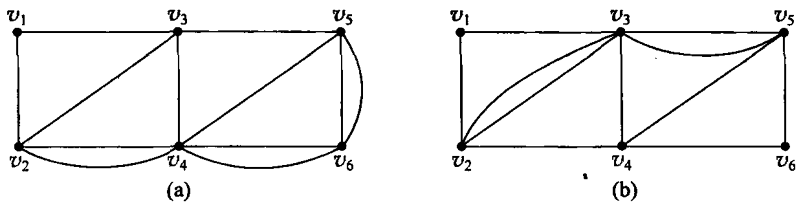


图 11-32

行方案称为最优方案。

现在的问题是第一个可行方案如何确定，在确定一个可行方案后，怎么判断这个方案是否为最优方案？若不是最优方案，如何调整这个方案？

1. 第一个可行方案的确定方法

在第 1 节中，我们已经证明，在任何一个图中，奇点个数必为偶数。所以如果图中有奇点，就可以把它们配成对。又因为图是连通的，故每一对奇点之间必有一条链，我们把这条链的所有边作为重复边加到图中去，可见新图中必无奇点，这就给出了第一个可行方案。

例 16 图 11-33 中的街道图，有四个奇点，v_2,v_4,v_6,v_8，将其分成两对，比如说 v_2 与 v_4 为一对，v_6 与 v_8 为一对。

在图 11-33 中，连接 v_2 与 v_4 的链有好几条，任取一条，例如取链($v_2,v_1,v_8,v_7,v_6,v_5,v_4$)。把边$[v_2,v_1]$，$[v_1,v_8]$，$[v_8,v_7]$，$[v_7,v_6]$，$[v_6,v_5]$，$[v_5,v_4]$作为重复边加到图中去，同样地取 v_6 与 v_8 之间的一条链($v_8,v_1,v_2,v_3,v_4,v_5,v_6$)，把边$[v_8,v_1]$，$[v_1,v_2]$，$[v_2,v_3]$，$[v_3,v_4]$，$[v_4,v_5]$，$[v_5,v_6]$也作为重复边加到图中去，于是得图 11-34。

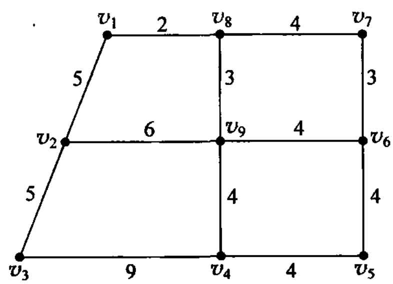


图 11-33

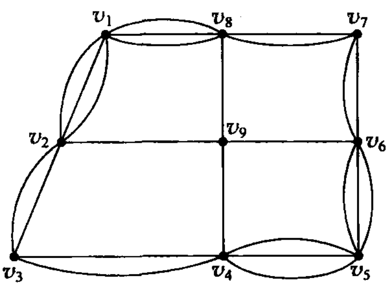


图 11-34

在图 11-34 中，没有奇点，对应于这个可行方案，重复边总权为：

$$2w_{12}+w_{23}+w_{34}+2w_{45}+2w_{56}+w_{67}+w_{78}+2w_{18}=51$$

2. 调整可行方案，使重复边总权下降

首先，从图 11-34 中可以看出，在边$[v_1,v_2]$上有两条重复边，如果把它们都从图中去掉，图仍然无奇点。即剩下的重复边还是一个可行方案，而总长度却有所下降。同样道

理，$[v_1,v_8]$，$[v_4,v_5]$，$[v_5,v_6]$上的重复边也是如此。

一般情况下，若边$[v_i,v_j]$上有两条或两条以上的重复边时，从中去掉偶数条，就能得到一个总权较小的可行方案。

因而有

(1) 在最优方案中，图的每一边上最多有一条重复边。

依此，图 11-34 可以调整为图 11-35，重复边总权下降为 21。

其次，我们还可以看到，如果把图中某个圈上的重复边去掉，而给原来没有重复边的边加上重复边，图中仍没有奇点。因而如果在某个圈上重复边的总权大于这个圈的总权的一半，像上面所说的那样作一次调整，将会得到一个总权下降的可行方案。

(2) 在最优方案中，图中每个圈上的重复边的总权不大于该圈总权的一半。

如在图 11-35 中，圈(v_2,v_3,v_4,v_9,v_2)的总权为 24，但圈上重复边总权为 14，大于该圈总权的一半。因此可以作一次调整，以$[v_2,v_9]$，$[v_9,v_4]$上的重复边代替$[v_2,v_3]$，$[v_3,v_4]$上的重复边，使重复边总权下降为 17，如图 11-36。

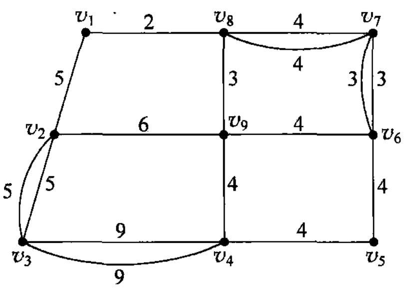


图 11-35

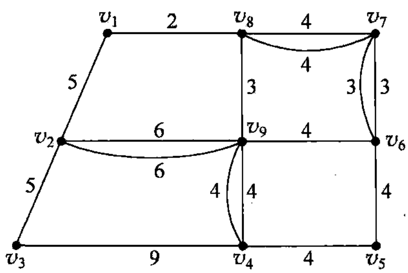


图 11-36

3. 判断最优方案的标准

从上面的分析中可知，一个最优方案一定是满足条件(1)和(2)的可行方案，反之，可以证明一个可行方案若满足条件(1)和(2)，则这个可行方案一定是最优方案。根据这样的判断标准，对给定的可行方案，检查它是否满足条件(1)和(2)。若满足，所得方案即为最优方案；若不满足，则对方案进行调整，直至条件(1)和(2)均得到满足时为止。

检查图 11-36 中的圈$(v_1,v_2,v_9,v_6,v_7,v_8,v_1)$，它的重复边总权为 13，而圈的总权为 24，不满足条件(2)，经调整得图 11-37。重复边总权下降为 15。

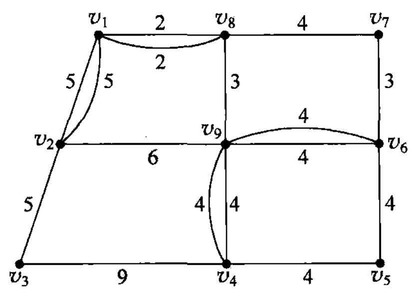


图 11-37

检查图 11-37，条件(1)和(2)均满足。于是得最优方案，图 11-37 中的任一个欧拉圈就是邮递员的最优邮递路线。

以上所说的求最优邮递路线的方法，通常称为奇偶点图上作业法。值得注意的是，方法的主要困

难在于检查条件(2)，它要求检查每一个圈。当图中点、边数较多时，圈的个数将会很多。如"日"字形图就有三个圈，而"田"字形的图就有 13 个圈。关于中国邮递员问题，已有比较好的算法，我们不去介绍它了。

注　记

希望进一步了解图的基本理论、图与网络的优化问题的算法和应用的读者，可参考参考资料[1]、[2]、[5]、[6]、[7]、[9]、[12]、[13]、[21]、[22]、[23]。参考资料[4]、[8]是两本供数学系和计算机系专业的研究生使用的教材，较全面地介绍图论及其应用中的基本概念、基本理论，以及近期的研究方向进展的问题。文献[11]是一本关于网络流理论和应用的很好入门书。在文献[1]、[9]、[14]中，作者们介绍了 Edmonds-Karp、Dinitz、Karzanov 等人给出的求最大流的改进算法。关于最小费用流问题，Tardos 等人相继提出了更好的算法，可参考文献[3]，该书全面系统地介绍了网络流及与之有关的若干图上的优化问题的理论、算法和应用。文献[15]是从实用的角度介绍网络流的概念、算法、计算程序及应用例子，涉及带增益函数的广义网络流、凸与凹费用函数下的最小费用流等。可供实际工作者参考。

中国邮递员问题的改进算法是 Edmonds 给出的。其算法涉及图的对集(matching)问题。文献[19]是一本关于对集理论的专著，详尽地介绍了对集及其推广的概念、重要成果、算法以及对集与网络流、线性规划、拟阵等分支的联系。文献[10]是一本关于欧拉图与有关论题的专著。

网络优化是一类组合优化问题。关于一般组合最优化的基本理论、算法及算法分析，可参考文献[16]、[17]、[20]。网络优化中，一个有名的问题是所谓旅行推销员问题(Traveling Salesman Problem)(也称为货郎担问题)：一个推销员要到若干个城市推销产品，然后回到出发点。已知每两个城市之间的距离，他应如何选择其旅行路线，使每个城市经过一次且仅仅一次，并且总的行程最短？表面上看，这个问题与中国邮递员问题很相似，但实际上，这是一个非常困难的问题，至今没有一个求最优旅行路线的有效方法。由 Lawler 等人编著的文献[18]全面地介绍了这个问题的历史、概况、理论发展和算法分析。

习题

11.1　证明如下序列不可能是某个简单图的次的序列：

(1) 7,6,5,4,3,2。

(2) 6,6,5,4,3,2,1。

(3) 6,5,5,4,3,2,1。

11.2　已知九个人 $v_1,v_2,\cdots,v_9$ 中，v_1 和两人握过手，v_2,v_3 各和四个人握过手，v_4,v_5,v_6,v_7 各和五个人握过手，v_8,v_9 各和六个人握过手，证明这九个人中一定可以找出三个人互相握过手。

11.3　用破圈法和避圈法找出图 11-38 的一个支撑树。

11.4　用破圈法和避圈法求图 11-39 中各图的最小树。

11.5　已知世界六大城市：(Pe)，(N)，(Pa)，(L)，(T)，(M)。试在由表 11-4所示交通网络的数据中确定最小树。

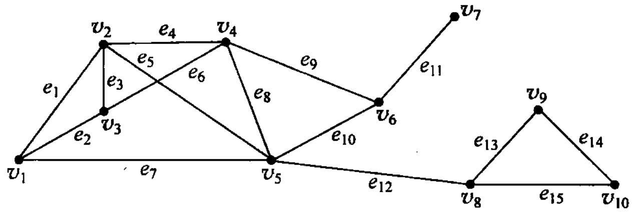


图 11-38

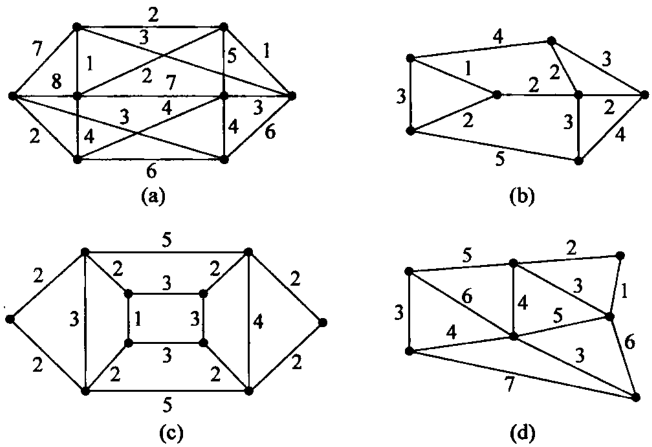


图 11-39

表 11-4

城市	Pe	T	Pa	M	N	L
Pe	×	13	51	77	68	50
T	13	×	60	70	67	59
Pa	51	60	×	57	36	2
M	77	70	57	×	20	55
N	68	67	36	20	×	34
L	50	59	2	55	34	×

11.6 有九个城市 $v_1, v_2, \cdots, v_9$，其公路网如图 11-40 所示。弧旁数字是该段公路的长度，有一批货物从 v_1 运到 v_9，问走哪条路最短？

11.7 用 Dijkstra 方法求图 11-41 中从 v_1 到各点的最短路。

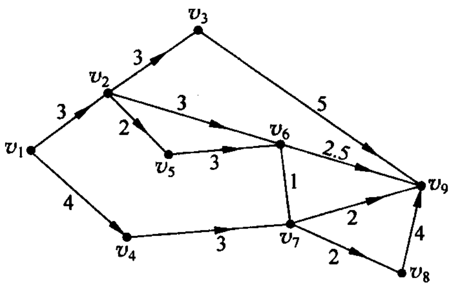


图 11-40

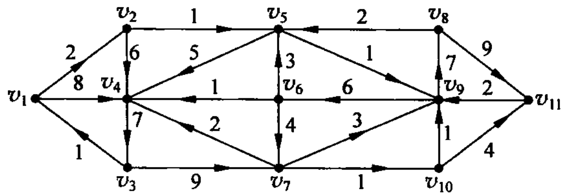


图 11-41

11.8 求图 11-42 中从 v_1 到各点的最短路。

11.9 在图 11-43 中

(1) 用 Dijkstra 方法求从 v_1 到各点的最短路；

(2) 指出对 v_1 来说哪些顶点是不可到达的。

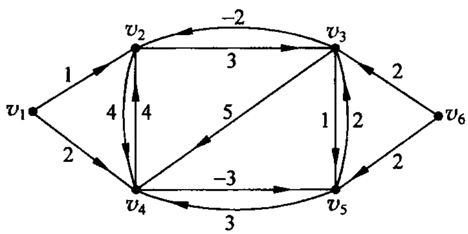


图 11-42

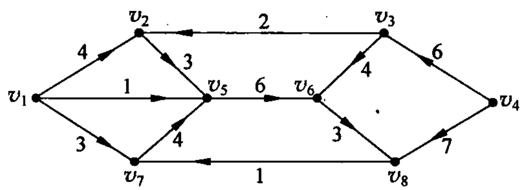


图 11-43

11.10 求图 11-44 中从任意一点到另外任意一点的最短路。

11.11 在如图 11-45 所示的网络中，每弧旁的数字是 (c_{ij}, f_{ij})。

(1) 确定所有的截集；

(2) 求最小截集的容量；

(3) 证明指出的流是最大流。

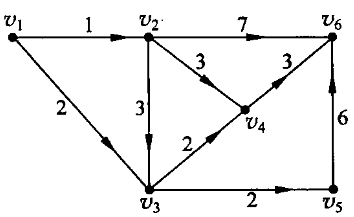


图 11-44

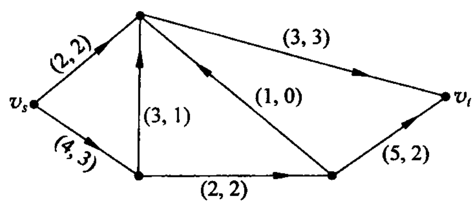


图 11-45

11.12 求如图 11-46 所示的网络的最大流(每弧旁的数字是 (c_{ij}, f_{ij}))。

11.13 求如图 11-47 所示的网络的最大流。

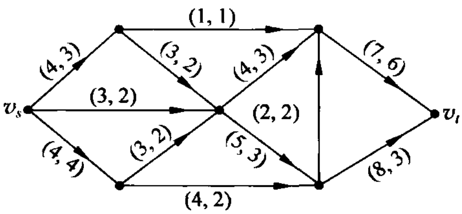


图 11-46

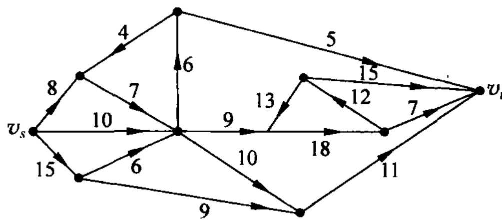


图 11-47

11.14 两家工厂 x_1 和 x_2 生产一种商品，商品通过如图 11-48 所示的网络运送到市场 y_1,y_2,y_3，试用标号法确定从工厂到市场所能运送最大总量。

11.15 求如图 11-49 所示的网络的最小费用最大流，每弧旁的数字是 (b_{ij},c_{ij})。

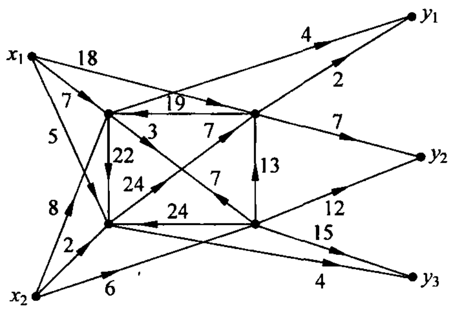


图 11-48

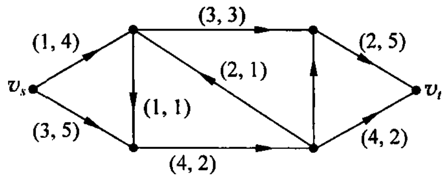


图 11-49

11.16 求解如图 11-50 所示的中国邮递员问题。

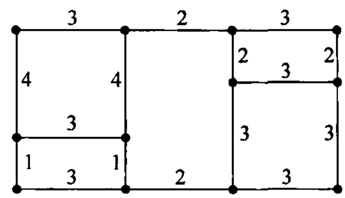


图 11-50

11.17 设 $G=(V,E)$ 是一个简单图，令 $\delta(G)=\min\limits_{v\in V}\{d(v)\}$（称 $\delta(G)$ 为 G 的最小次）。证明：

(1) 若 $\delta(G)\geqslant 2$，则 G 必有圈；

(2) 若 $\delta(G)\geqslant 2$，则 G 必有包含至少 $\delta(G)+1$ 条边的圈。

11.18 设 G 是一个连通图，不含奇点。证明：G 中不含割边。

11.19 给一个连通赋权图 G，类似于求 G 的最小支撑树的 Kruskal 方法，给出一个求 G 的最大支撑树的方法。

11.20 下述论断正确与否：可行流 f 的流量为零，即 $v(f)=0$，当且仅当 f 是零流。

11.21 设 $D=(V,A,C)$ 是一个网络。证明：如果 D 中所有弧的容量 c_{ij} 都是整数，那么必存在一个最大流 $f=\{f_{ij}\}$，使所有 f_{ij} 都是整数。

11.22 已知有六台机床 $x_1,x_2,\cdots,x_6$，六个零件 $y_1,y_2,\cdots,y_6$。机床 x_1 可加工零件 y_1；x_2 可加工零件 y_1,y_2；x_3 可加工零件 y_1,y_2,y_3；x_4 可加工零件 y_2；x_5 可加工零件 y_2,y_3,y_4；x_6 可加工零件 y_2,y_5,y_6。现在要求制定一个加工方案，使一台机床只加工一

个零件，一个零件只在一台机床上加工，要求尽可能多地安排零件的加工。试把这个问题化为求网络最大流的问题，求出能满足上述条件的加工方案。

参考资料

[1] 田丰，马仲蕃. 图与网络流理论. 北京：科学出版社，1987

[2] 谢金星，邢文顺. 网络优化. 北京：清华大学出版社，2000

[3] Ahuja R K, Magnanti T L & Orlin J B. *Network Flows Theory Algorithms and Applications*. Prentice-Hall, 1993

[4] Bollobas B. *Modern Graph Theory*. Grad Texts Math. 184, Springer, 1998

[5] Bondy J A & Murty U S R. *Graph Theory with Applications*. The Macmillan Press, 1976
吴望名，李念祖等译. 图论及其应用. 北京：科学出版社，1984

[6] Chartrand G & Oellermann O R. *Applied and Algorithmic Graph Theory*. McGraw-Hill, 1993

[7] Deo N. *Graph Theory with Applications in Engineering and Computer Science*. Prentice-Hall, 1974

[8] Diestel R. *Graph Theory*. Grad. Texts Math. 173, Springer-Verlag, 2000

[9] Even S. *Graph Algorithms*. Computer Science Press, 1979

[10] Fleischner H. *Eulerian Graphs and Related Topics*. Ann. Dis. Math. 45, North Holland, Amsterdam, 1990

[11] Ford L R & Fulkerson D R. *Flows in Networks*. Princeton University Press, 1962

[12] Foulds L R. *Graph Theory Applications*. Springer-Velag, 1992

[13] Gibbons A. *Algorithmic Graph Theory*. Cambridge University Press, 1985

[14] Hu T C. *Combinatorial Algorithms*. Addison-Wesley Publishing Company, 1982

[15] Jensen P A & Barnes J W. *Network Flow Programming*. John Wiely & Sons, 1980
孙东川译. 网络流规划. 北京：科学出版社，1988

[16] Korte B & Vygen J. *Combinatorial Optimization. Theory and Algorithms*. Springer, 1991

[17] Lawler E L. *Combinatorial Optimization: Networks and Matroids*. Holt Rinehart and Winston, 1976

[18] Lawler E L, Lenstra J K & Rinooy-Kan A H G. *The Traveling Salesman Problem*. Wiley-Interscience. John Wiley & Sons, 1985

[19] Lovasz L & Plummer M D. *Matching Theory*. Elsevier Science Publishing Company Inc, 1986

[20] Papadimitriou C H & Steiglitz K. *Combinatorial Optimization. Algorithms and Complexity*. Prentice-Hall, 1982
刘振宏，蔡茂诚译. 组合最优化：算法和复杂性. 北京：清华大学出版社，1988

[21] Swamy M N S & Thulasiraman K. *Graphs Networks and Algorithms*. Wiley-Interscience, John Wiley & Sons, 1981

[22] West D B. *Introduction to Graph Theory*. Prentice-Hall, 1993

[23] Wilson R J & Beineke W L. *Applications of Graph Theory*. Academic Press, 1979

CHAPTER 12 第12章

网络计划

当前，世界上工业发达国家都非常重视网络计划技术在现代管理中的应用。它已被许多国家公认为当前最为行之有效的管理方法之一。国外多年实践证明，应用网络计划技术组织与管理生产和项目一般能缩短工期20%左右，降低成本10%左右。

美国是网络计划技术的发源地。美国政府于1962年规定，凡与政府签订合同的企业，都必须采用网络计划技术，以保证工程进度和质量。1974年麻省理工学院调查指出："绝大部分美国公司采用网络计划编制施工计划。"目前，美国基本上实现了用计算机绘画、优化计算和资源平衡、项目进度控制，实现了计划工作自动化。以后又提出了新的网络计划技术，如图示评审技术(GERT)，风险评审技术(VERT)等。

我国应用网络计划技术是从20世纪60年代初期开始。著名科学家钱学森将网络计划方法引入我国，并在航天系统应用。著名数学家华罗庚在综合研究各类网络方法的基础上，结合我国实际情况加以简化，于1965年发表了《统筹方法平话》，为推广应用网络计划方法奠定了基础。近几年，随着科技的发展和进步，网络计划技术的应用也日趋得到工程管理人员的重视，且已取得可观的经济效益。如上海宝钢炼铁厂1号高炉土建工程施工中，应用网络法，缩短工期21%，降低成本9.8%。广州白天鹅宾馆在建设中，运用网络计划技术，工期比外商签订的合同提前四个半月，仅投资利息就节约1 000万港元。为在我国推广普及网络计划技术，我国建设部公布了《工程网络计划技术规程》，以便统一技术术语、符号、代号和计算规范。特别近几年来，微机的普及和网络计划软件的不断更新换代。这些为在国内大范围推广网络计划技术创造了条件。

12.1 网络计划图

网络计划图的基本思想是，首先应用网络计划图来表示工程项目中计划要完成的各项工作，完成各项工作必然存在先后顺序及其相互依赖的逻辑关系；这些关系用节点、箭线来构成网络图。网络图是由左向右绘制，表示工作进程，并标注工作名称、代号和工作持续时间等必要信息。通过对网络计划图进行时间参数的计算，找出计划中的关键工作和关键线路；通过不断改进网络计划，寻求最优方案，以求在计划执行过程中对计划进行有效的控制与监督，保证合理地使用人力、物力和财力，以最小的消耗取得最大的经济

效果。

12.1.1 基本术语

网络计划图是在网络图上标注时标和时间参数的进度计划图，实质上是有时序的有向赋权图。表述关键路线法（CPM）和计划评审技术（PERT）的网络计划图没有本质的区别，它们的结构和术语是一样的。仅前者的时间参数是确定型的，而后者的时间参数是不确定型的。于是统一给出一套专用的术语和符号。

（1）节点、箭线是网络计划图的基本组成元素。箭线是一线段带箭头的实射线（用“→”表示）和虚射线（用“-->”表示），节点是箭线两端的连接点（用“○”或“□”表示）。

（2）工作（也称工序、活动、作业），将整个项目按需要粗细程度分解成若干需要耗费时间或需要耗费其他资源的子项目或单元。它们是网络计划图的基本组成部分。

（3）描述工程项目网络计划图有两种表达的方式：双代号网络计划图和单代号网络计划图。双代号网络计划图在计算时间参数时，又可分为工作计算法和节点计算法。

（4）双代号网络计划图。在双代号网络计划图中，用箭线表示工作，箭尾的节点表示工作的开始点，箭头的节点表示工作的完成点。用（i-j）两个代号及箭线表示一项工作，在箭线上标记必需的信息，如图 12-1 所示。

箭线之间的连接顺序表示工作之间的先后开工的逻辑关系。

（5）单代号网络计划图。用节点表示工作，箭线表示工作之间的先完成与后完成的关系为逻辑关系。在节点中标记必需的信息，如图 12-2 所示。

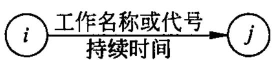


图 12-1

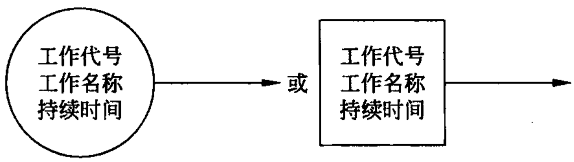


图 12-2

12.1.2 双代号网络计划图

这里主要介绍双代号网络计划图的绘制和按工作计算时间参数的方法。以下通过例题来说明网络计划图的绘制和时间参数的计算。

例 1 开发一个新产品，需要完成的工作和先后关系，各项工作需要的时间汇总在逻辑关系表中，见表 12-1。要求编制该项目的网络计划图和计算有关参数，根据表 12-1 中数据，绘制网络图，见图 12-3。

表 12-1

序号	工 作 名 称	工 作 代 号	工作持续时间/天	紧 后 工 作
1	产品设计和工艺设计	A	60	B,C,D,E
2	外购配套件	B	45	L
3	锻件准备	C	10	F
4	工装制造 1	D	20	G,H
5	铸件	E	40	H
6	机械加工 1	F	18	L
7	工装制造 2	G	30	K
8	机械加工 2	H	15	L
9	机械加工 3	K	25	L
10	装配与调试	L	35	/

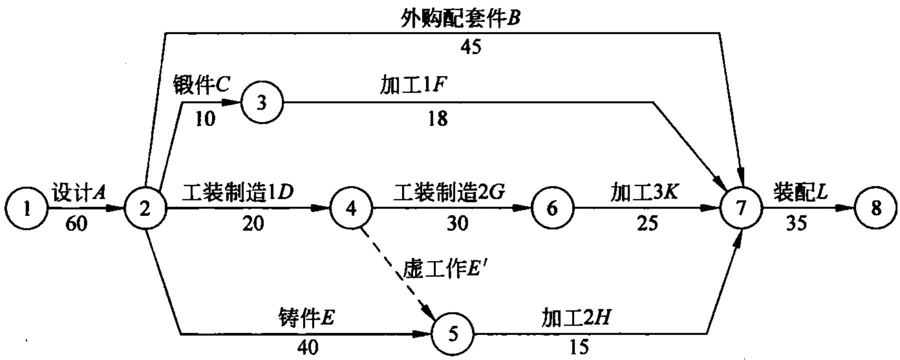


图 12-3

为了正确表述工程项目中各个工作的相互连接关系和正确绘制网络计划图，应遵循以下规则和术语：

1. 网络计划图的方向、时序和节点编号

网络计划图是有向、有序的赋权图，按项目的工作流程自左向右地绘制。在时序上反映完成各项工作的先后顺序。节点编号必须按箭尾节点的编号小于箭头节点的编号来标记。在网络图中只能有一个起始节点，表示工程项目的开始。一个终点节点，表示工程项目的完成。从起始节点开始沿箭线方向顺序自左往右，通过一系列箭线和节点，最后到达终点节点的通路，称为线路。

2. 紧前工作和紧后工作

紧前工作是指紧排在本工作之前的工作，且开始或完成后，才能开始本工作。紧后工作是指紧排在本工作之后的工作，且本工作开始或完成后，才能做的工作。如图 12-3 中，只有工作 A 完成后工作 B,C,D,E 才能开始，工作 A 是 B,C,D,E 的紧前工作；而工作 B,C,D,E 则是工作 A 的紧后工作。在复杂的工程项目中，它们之间有三种关系：结束

后，才开始；开始后，才开始；结束后，才结束。本例只涉及结束后才开始的关系。从起始节点至本工作之前在同一线路的所有工作，称为先行工作；自本工作到终点节点在同一线路的所有工作，称为后继工作。工作 G 的先行工作有工作 A，D；工作 K，L 是工作 G 的后继工作。

3. 虚工作

在双代号网络计划图中，只表示相邻工作之间的逻辑关系，不占用时间和不消耗人力、资金等的虚设的工作。虚工作用虚箭线 ----→ 表示。如在图 12-3 中④----→⑤只表示工作 D 完成后，工作 H 才能开始。

4. 相邻两节点之间只能有一条箭线连接，否则将造成逻辑上的混乱

如图 12-4 是错误的画法，为了使两节点之间只有一条箭线，可增加一个节点②′，并增加一项虚工作②′⋯→②。图 12-5 是正确的画法。

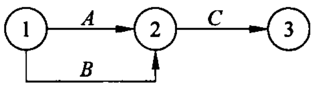


图 12-4

应当改成

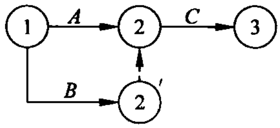


图 12-5

5. 网络计划图中不能有缺口和回路

在网络计划图中严禁出现从一个节点出发，顺箭线方向又回到原出发节点，形成回路。回路将表示这工作永远不能完成。网络计划图中出现缺口，表示这些工作永远达不到终点。项目无法完成。

6. 平行工作

可与本工作同时进行的工作。

7. 起始节点与终点节点

在网络计划图中只能有一个起始节点和一个终点节点。当工程开始或完成时存在几个平行工作时，可以用虚工作将它们与起始节点或终点节点连接起来。

8. 线路

网络图中从起点节点沿箭线方向顺序通过一系列箭线与节点，最后到达终点节点的通路。本例中有五条线路。并可以计算出各线路的持续时间，见表 12-2。

表 12-2

线路	线路的组成	各工作的持续时间之和(天)
1	①→②→⑦→⑧	60＋45＋35＝140
2	①→②→③→⑦→⑧	60＋10＋18＋35＝123
3	①→②→④→⑥→⑦→⑧	60＋20＋30＋25＋35＝170
4	①→②→④→⑤→⑦→⑧	60＋20＋15＋35＝130
5	①→②→⑤→⑦→⑧	60＋40＋15＋35＝150

从网络图中可以计算出各线路的持续时间。其中有一条线路的持续时间最长线路是关键路线，或称为主要矛盾线。关键路线上的各工作为关键工作。因为它的持续时间就决定了整个项目的工期。关键路线的特征以后再进一步阐述。

9. 网络计划图的布局

尽可能将关键路线布置在网络计划图的中心位置，按工作的先后顺序将联系紧密的工作布置在邻近的位置。为了便于在网络计划图上标注时间等数据，箭线应是水平线或具有一段水平线的折线。在网络计划图上附有时间坐标或日历进程。

10. 网络计划图的类型

(1) 总网络计划图，以整个项目为计划对象，编制网络计划图。供决策领导层使用。

(2) 分级网络计划图，这是按不同管理层次的需要，编制的范围大小不同，详细程度不同的网络计划图，供不同管理部门使用。

(3) 局部网络计划图，将整个项目某部分为对象，编制更详细的网络计划图，供专业部门使用。

当用计算机网络计划软件编制时，在计算机上可进行网络计划图分解与合并。网络计划图详细程度，可以根据需要，将工作分解为更细的子工作，也可以将几项工作合并为综合的工作，以便显示不同粗细程度的网络计划。

12.2 网络计划图的时间参数计算

网络计划的时间参数计算有几种类型：双代号网络计划有工作计算法和节点计算法；单代号网络计划有节点计算法。以下仅介绍工作计算法，其他的计算法可参考《工程网络计划技术规程》(JGJ/T121-99)。

网络图中工作的时间参数，它们是：工作持续时间(D)，工作最早开始时间(ES)，工作最早完成时间(EF)；工作最迟开始时间(LS)，工作最迟完成时间(LF)；工作总时差(TF)和工作自由时差(FF)。

12.2.1 工作持续时间 D

工作持续时间计算是一项基础工作，关系到网络计划是否能得到正确实施。为了有效地使用网络计划技术，需要建立相应的数据库。这需要专项讨论的问题。这里简述计算工作持续时间的两类数据和两种方法。

1. 单时估计法(定额法)

每项工作只估计或规定一个确定的持续时间值的方法。一般有工作的工作量，劳动定额资料以及投入人力的多少等，计算各工作的持续时间

$$D=\frac{Q}{R\cdot S\cdot n}$$

Q——工作的工作量。以时间单位表示，如小时，或以体积、质量、长度等单位表示；

R——可投入人力和设备的数量；

S——每人或每台设备每工作班能完成的工作量；

n——每天正常工作班数。

当具有类似工作的持续时间的历史统计资料时，可以根据这些资料，采用分析对比的方法确定所需工作的持续时间。

2. 三时估计法

在不具备有关工作的持续时间的历史资料时，在较难估计出工作持续时间时，可对工作进行估计三种时间值，然后计算其平均值。这三种时间值是：

乐观时间——在一切都顺利时，完成工作需要的最少时间，记作 a。

最可能时间——在正常条件下，完成工作所需要的时间，记作 m。

悲观时间——在不顺利条件下，完成工作需要最多的时间，记作 b。

显然上述三种时间发生都具有一定的概率，根据经验，这些时间的概率分布认为是正态分布。一般情况下，通过专家估计法，给出三时估计的数据。可以认为工作进行时出现最顺利和最不顺利的情况比较少，较多是出现正常的情况。按平均意义可用以下公式计算工作持续时间值：$D=\frac{a+4m+b}{6}$；方差 $\sigma^2=\left(\frac{b-a}{6}\right)^2$。

12.2.2 计算关系式

这些时间参数的关系可以用图 12-6 表示工作的关系状态。手工计算可在网络图上进行，计算步骤为：

(1) 计算各路线的持续时间(见表 12-2)。

(2) 按网络图的箭线的方向，从起始工作开始，计算各工作的 ES，EF。

(3) 从网络图的终点节点开始，按逆箭线的方向，推算出各工作的 LS，LF。

(4) 确定关键路线(CP)。

(5) 计算 TF，FF。

(6) 平衡资源。

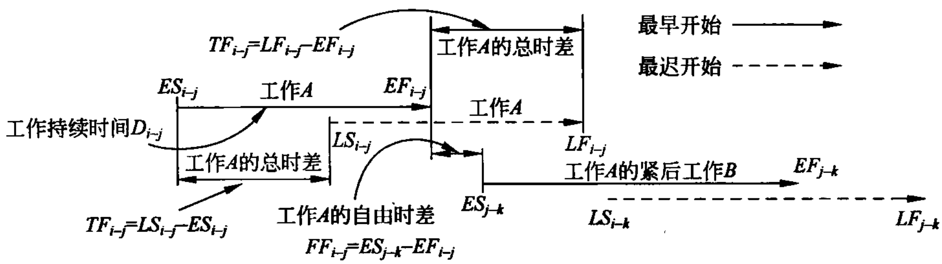


图 12-6

例 2 计算各工作的时间参数，并将计算结果记入网络计划图的相应工作的□中，见图 12-7。

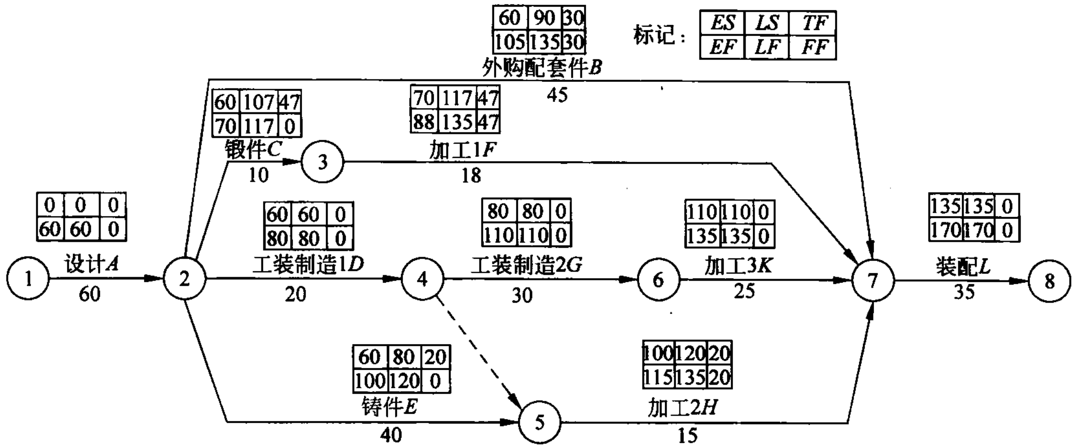


图 12-7

1. 工作最早开始时间 ES 和工作最早完成时间 EF 的计算

利用网络计划图，从网络计划图的起始点开始，沿箭线方向依次逐项计算。第一项工作的最早开始时间为 0，记作 $ES_{i-j}=0$。(起始点 $i=1$)。第一件工作的最早完成时间 $EF_{1-j}=ES_{1-j}+D_{1-j}$。第一件工作完成后，其紧后工作才能开始。它工作最早完成时间 EF 就是其紧后工作最早开始时间 ES。本工作的持续时间 D。表示为

$$EF_{i-j}=ES_{i-j}+D_{i-j}。$$

计算工作的 ES 时，当有多项紧前工作情况下，只能在这些紧前工作中都完成后才能开始。因此本工作的最早开始时间是：$ES=\max$(紧前工作的 EF)其中 $EF=ES+$工作持续时间 D，表示为

$$ES_{i-j}=\max_h(EF_{h-i})=\max_h(ES_{h-i}+D_{h-i}),$$

例 2 的 ES,EF 计算值在表 12-3 的③，④列中。

表 12-3

工作 $i-j$	持续时间 D_{i-j}	最早开始时间 ES_{i-j}	最早完成时间 EF_{i-j}
①	②	③	④=③+②
$A(1-2)$	60	$ES_{1-2}=0$	$EF_{1-2}=ES_{1-2}+D_{1-2}=0+60=60$
$B(2-7)$	45	$ES_{2-7}=EF_{1-2}=60$	$EF_{2-7}=ES_{2-7}+D_{2-7}=60+45=105$
$C(2-3)$	10	$ES_{2-3}=EF_{1-2}=60$	$EF_{2-3}=ES_{2-3}+D_{2-3}=60+10=70$
$D(2-4)$	20	$ES_{2-4}=EF_{1-2}=60$	$EF_{2-4}=ES_{2-4}+D_{2-4}=60+20=80$
$E(2-5)$	40	$ES_{2-5}=EF_{1-2}=60$	$EF_{2-5}=ES_{2-5}+D_{2-5}=60+40=100$
$E'(4-5)$	0(虚工作)	$ES_{4-5}=EF_{2-4}=80$	$EF_{4-5}=ES_{4-5}+D_{4-5}=80+0=80$
$F(3-7)$	18	$ES_{3-7}=EF_{2-3}=70$	$EF_{3-7}=ES_{3-7}+D_{3-7}=70+18=88$
$G(4-6)$	30	$ES_{4-6}=EF_{2-4}=80$	$EF_{4-6}=ES_{4-6}+D_{4-6}=80+30=110$

续表

工作 $i-j$	持续时间 D_{i-j}	最早开始时间 ES_{i-j}	最早完成时间 EF_{i-j}
①	②	③	④=③+②
$H(5-7)$	15	$ES_{5-7}=\max(EF_{2-5},EF_{4-5})$ $=EF_{2-5}=100$	$EF_{5-7}=ES_{5-7}+D_{5-7}=100+15=115$
$K(6-7)$	25	$ES_{6-7}=EF_{4-6}=110$	$EF_{6-7}=ES_{6-7}+D_{6-7}=110+25=135$
$L(7-8)$	35	$ES_{7-8}=\max(EF_{2-7},EF_{3-7},EF_{6-7},$ $EF_{5-7})=EF_{6-7}=135$	$EF_{7-8}=ES_{7-8}+D_{7-8}=135+35=170$

利用双代号的特征，很容易在表中确定某工作的紧前工作和紧后工作。凡是后续工作的箭尾代号与某工作的箭头代号相同者，便是它的紧后工作；凡是先行工作的箭头代号与某工作的箭尾代号相同者，便是它的紧前工作。在表 12-3 中首先填入①、②两列数据，然后由上往下计算 ES 与 EF。若某工作$(i-j)$的先行工作中存在几个$(h-i)$，从中选择最大的 EF_{h-i}进行计算 $ES_{i-j}=\max_h[EF_{h-i}]$，即计算 EF_{i-j}，如计算 ES_{7-8}时，可从表 12-3的④列已有的 EF_{6-7}，EF_{5-7}，EF_{3-7}中找到最大的 $EF_{6-7}=135$。将它填入表 12-3的③列，对应的 $L(7-8)$行即可。如此计算也很方便。

2. 工作最迟开始时间 *LS* 与工作最迟完成时间 *LF*

应从网络图的终点节点开始，采用逆序法逐项计算。即按逆箭线方向，依次计算各工作的最迟完成时间 LF 和最迟开始时间 LS，直到第一项工作为止。网络图中最后一项工作$(i-n)(j=n)$的最迟完成时间应由工程的计划工期确定。在未给定时，可令其等于其最早完成时间，即 $LF_{i-n}=EF_{i-n}$。EF_{i-n}由表 12-3 中的计算结果是已知的了，并且应当小于或等于计划工期规定的时间 T_r。

$$LF=\min(\text{紧后工作的 }LS),LS=LF-\text{工作持续时间 }D$$

其他工作的最迟开始时间 $LS_{i-j}=LF_{i-j}-D_{i-j}$；当有多个紧后工作时，最迟完成时间 $LF=\min$(紧后工作的 LS)，或表示为 $LF_{i-j}=\min_k(LF_{j-k}-D_{j-k})$。

可在表 12-4 中进行。计算从下到上地进行，从工作 $L(7-8)$开始，令表 12-4 的⑤列最后一行 $LF_{7-8}=EF_{7-8}=170$。

表 12-4

工作 $i-j$	持续时间 D_{i-j}	最迟完成时间 $LF_{i-j}=\min_k(LS_{j-k})$	最迟开始时间 $LS_{i-j}=LF_{i-j}-D_{i-j}$	总时差 TF_{i-j} $=LS_{i-j}-ES_{i-j}$	自由时差 $FF_{i-j}=ES_{i-k}-EF_{i-j}$
①	②	⑤	⑥=⑤-②	⑦=⑥-③	⑧
$A(1-2)$	60	$LF_{1-2}=LS_{2-4}=60$	$LS_{1-2}=LF_{1-2}-60=60-60=0$	$0-0=0$	$FF_{1-2}=ES_{2-3}-EF_{1-2}=0$
$B(2-7)$	45	$LF_{2-7}=LS_{7-8}=135$	$LS_{2-7}=LF_{2-7}-45=135-45=90$	$90-60=30$	$FF_{2-7}=ES_{7-8}-EF_{2-7}=135-105=30$
$C(2-3)$	10	$LF_{2-3}=LS_{3-7}=117$	$LS_{2-3}=LF_{2-3}-10=117-10=107$	$107-60=47$	$FF_{2-3}=ES_{3-7}-EF_{2-3}=70-70=0$
$D(2-4)$	20	$LF_{2-4}=LS_{4-6}=80$	$LS_{2-4}=LF_{2-4}-20=80-20=60$	$60-60=0$	$FF_{2-4}=ES_{4-6}-EF_{2-6}=80-80=0$
$E(2-5)$	40	$LF_{2-5}=LS_{5-7}=120$	$LS_{2-5}=LF_{2-5}-40=120-40=80$	$80-60=20$	$FF_{2-5}=ES_{5-7}-EF_{2-5}=100-100=0$
$F(3-7)$	18	$LF_{3-7}=LS_{7-8}=135$	$LS_{3-7}=LF_{3-7}-18=135-18=117$	$117-70=47$	$FF_{3-7}=ES_{7-8}-EF_{3-7}=135-88=47$
$G(4-6)$	30	$LF_{4-6}=LS_{6-7}=110$	$LS_{4-6}=LF_{4-6}-30=110-30=80$	$80-80=0$	$FF_{4-6}=ES_{6-7}-EF_{4-6}=110-110=0$
$H(5-7)$	15	$LF_{5-7}=LS_{7-8}=135$	$LS_{5-7}=LF_{5-7}-15=135-15=120$	$120-100=20$	$FF_{5-7}=ES_{7-8}-EF_{5-7}=135-115=20$
$K(6-7)$	25	$LF_{6-7}=LS_{7-8}=135$	$LS_{6-7}=LF_{6-7}-25=135-25=110$	$110-110=0$	$FF_{6-7}=ES_{7-8}-EF_{6-7}=135-135=0$
$L(7-8)$	35	$LF_{7-8}=EF_{7-8}=170$	$LS_{7-8}=LF_{7-8}-35=170-35=135$	$135-135=0$	$FF_{7-8}=T-170=170-170=0$

于是可计算出 $EF_{7-8}=LF_{7-8}-D_{7-8}=135$。工作 $L(7-8)$的紧前工作的箭尾代号与工作 $L(7-8)$的箭头代号是相同的，这里有 $K(6-7)$，$H(5-7)$，$F(3-7)$，$B(2-7)$；它们只有唯一的紧后工作 $L(7-8)$，所以 LF_{6-7}，LF_{5-7}，LF_{3-7}，LF_{2-7}都等于 $LF_{7-8}=135$。填入表 12-4⑤列的相应行即可。当具有多个紧后工作时，如要计算 LF_{1-2}时，先查 $A(1-2)$的紧后工作有几个，从代号可以看到是 $B(2-7)$，$C(2-3)$，$D(2-4)$，$E(2-5)$，对应的有 $LS_{2-7}=90$，$LS_{2-3}=107$，$LS_{2-4}=60$，$LS_{2-5}=80$。其中最小的是 60，即 $LF_{1-2}=LS_{2-4}=60$。

3. 工作时差

工作时差是指工作有机动时间。常用有两种时差，即工作总时差和工作自由时差。

(1) 工作总时差 TF_{i-j}

TF_{i-j}是指在不影响工期的前提下，工作所具有的机动时间，按工作计算法计算。在表 12-4 中⑦＝⑥－③的数据。

$$TF_{i-j}=EF_{i-j}-ES_{i-j}-D_{i-j}=LS_{i-j}-ES_{i-j} \text{或} TF_{i-j}=LF_{i-j}-EF_{i-j}$$

注 工作总时差往往为若干项工作共同拥有的机动时间，如工作 $C(2-3)$和工作 $F(3-7)$，其工作总时差为 47，当工作 $C(2-3)$用去一部分机动时间后，工作 $F(3-7)$的机动时间将相应地减少。

(2) 工作自由时差 FF

工作自由时差是指：在不影响其紧后工作最早开始的前提下，工作所具有的机动时间。

$$FF_{i-j}=ES_{j-k}-ES_{i-j}-D_{i-j}\text{；或 } FF_{i-j}=ES_{j-k}-EF_{i-j}$$

计算结果见表 12-4⑧和图 12-7。工作自由时差是某项工作单独拥有的机动时间，其大小不受其他工作机动时间的影响。

关键路线的特征：在线路上从起点到终点都由关键工作组成。在确定型网络计划中是指线路中工作总持续时间最长的线路。在关键线路上无机动时间，工作总时差为零。在非确定型网络计划中是指估计工期完成可能性最小的线路。

12.3 时标网络计划图

时间坐标简称时标。时标在网络计划图的上方或下方，用以表示工程进度时间的坐标轴。根据需要规定时间单位为：小时、天、周、月或季。标注有时间坐标的网络计划图称为时标网络计划图。在该图中箭线的长度就表示工作持续时间的长度。在图中可以用实粗箭线表示关键工作和关键线路，并且可用不同的线型表示出工作的总时差和自由时差，例 1 的时标网络计划图如图 12-8 所示。

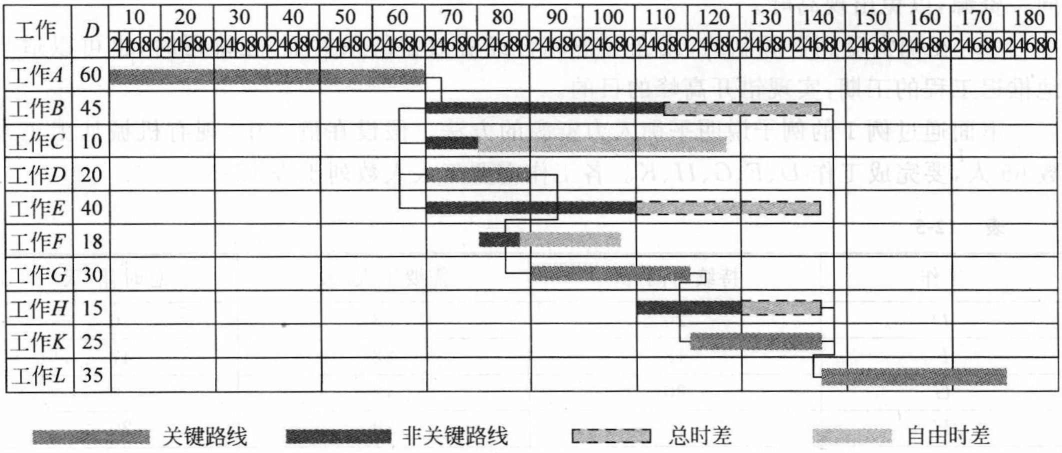


图 12-8

12.4 网络计划的优化

绘制网络计划图，计算时间参数和确定关键线路，仅得到一个初始计划方案。然后根据上级要求和实际资源的配置，需要对初始方案进行调整和完善，即进行网络计划优化。目标是综合考虑进度，合理利用资源，降低费用等。

12.4.1 工期优化

若网络计划图的计算工期大于上级要求的工期时，必须根据要求计划的进度，缩短工程项目的完工工期。主要采取以下措施，增加对关键工作的投入，以便缩短关键工作的持续时间，实现工期缩短。

(1) 采取技术措施，提高工效，缩短关键工作的持续时间，使关键线路的时间缩短。

(2) 采取组织措施，充分利用非关键工作的总时差，合理调配人力、物力和资金等资源。

12.4.2 资源优化

在编制初始网络计划图后，需要进一步考虑尽量利用现有资源的问题。即在项目的工期不变的条件下，均衡地利用资源。实际工程项目包括工作繁多，需要投入资源种类很多，均衡地利用资源是很麻烦的事，要用计算机来完成。为了简化计算，具体操作如下：

(1) 优先安排关键工作所需要的资源。

(2) 利用非关键工作的总时差，错开各工作的开始时间，避开在同一时区内集中使用

同一资源，以免出现高峰。

(3) 在确实受到资源制约，或在考虑综合经济效益的条件下，在许可时，也可以适当地推迟工程的工期，实现错开高峰的目的。

下面通过例 1 的例子说明平衡人力资源的方法。假设在例 1 中，现有机械加工工人数 65 人，要完成工作 D、F、G、H、K。各工作需要工人人数列于表 12-5。

表 12-5

工作	持续时间/天	需要工人/人	总时差/天
D	20	58	0
F	18	22	47
G	30	42	0
H	15	39	20
K	25	26	0

由于机械加工工人人数的限制。若上述工作都按最早开始时间安排，在完成各关键工作的 75 天工期中，每天需要机械加工工人人数如图 12-9 所示。有 10 天需要 80 人，另 10 天需要 81 人。超过了现有机械工人人数的约束，必须进行调整。以虚线表示的非关键路线上非关键工作 F、H 有机动时间，若将工作 F 延迟 10 天开工，就可以解决第 70～80天的超负荷问题；将工作 H 推迟 10 天开工，可以解决第 100～110 天的超负荷问题。于是新的负荷图(见图 12-10)能满足机械工人的人数 65 人约束条件。以上人力资源平衡是利用非关键工作的总时差，可以避开资源负荷的高峰。

避开资源负荷高峰时，可以采用将非关键工作分段作业或采用技术措施减少所需要的资源，也可以根据计划规定适当延长项目的工期。

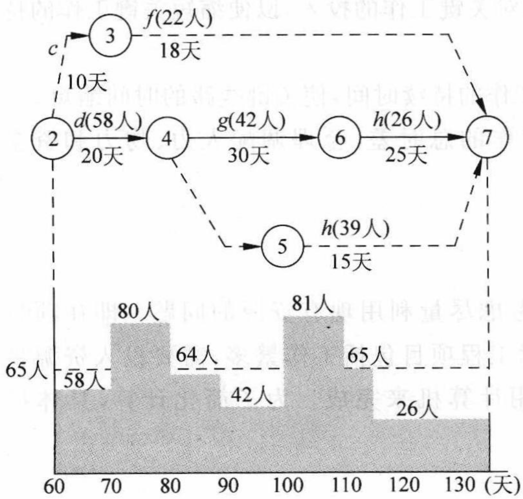


图 12-9

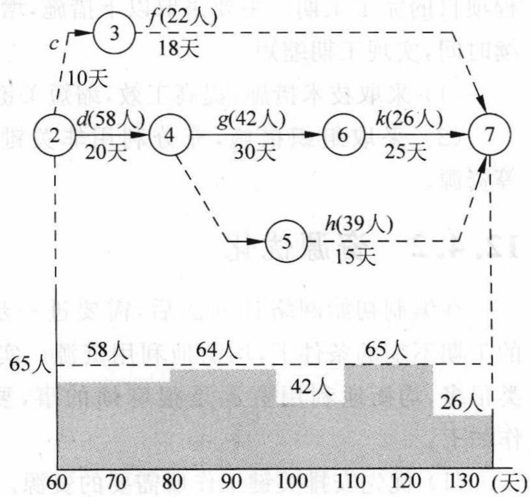


图 12-10

12.4.3 时间—费用优化

编制网络计划时，要研究如何使完成项目的工期尽可能缩短，费用尽可能少；或在保证既定项目完成时间条件下，所需要的费用最少；或在费用限制的条件下，项目完工的时间最短。这就是时间—费用优化要解决的问题。完成一项目的费用可以分为两大类：

1. 直接费用

直接与项目的规模有关的费用，包括材料费用，直接生产工人工资等。为了缩短工作的持续时间和工期，就需要增加投入，即增加直接费用。

2. 间接费用

间接费用包括管理费等。一般按项目工期长度进行分摊，工期愈短，分摊的间接费用就愈少。一般项目的总费用与直接费用、间接费用、项目工期之间存在一定关系，可以用图 12-11 表示。

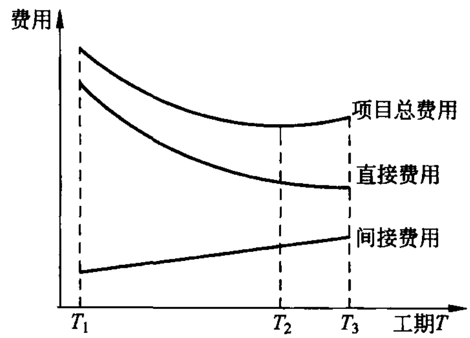


图 12-11 工期与总费用的关系曲线

图中，T_1——最短工期，项目总费用最高；

T_2——最佳工期；

T_3——正常的工期。

当总费用最少工期短于要求工期时，这就是最佳工期。

进行时间-费用优化时，首先要计算出不同工期下最低直接费用率，然后考虑相应的间接费用。费用优化的步骤如下：

(1) 计算工作费用增加率(简称费用率)

费用增加率是指缩短工作持续时间每一单位时间(如一天)所需要增加的费用。

按工作的正常持续时间计算各关键工作的费用率，通常可表示为

$$\Delta C_{i-j}=\frac{CC_{i-j}-CN_{i-j}}{DN_{i-j}-DC_{i-j}}$$

ΔC_{i-j}——工作 $i-j$ 的费用率；

CC_{i-j}——将工作 $i-j$ 持续时间缩短为最短持续时间后，完成该工作所需要的直接费用；

CN_{i-j}——在正常条件下完成工作 $i-j$ 所需要的直接费用；

DN_{i-j}——工作 $i-j$ 正常持续时间；

DC_{i-j}——工作 $i-j$ 最短持续时间。

(2) 在网络计划图找出费用率最低的一项关键工作或一组关键工作作为缩短持续时间的对象。其缩短后的值不能小于最短持续时间，不能成为非关键工作。

(3) 同时计算相应的增加的总费用，然后考虑由于工期的缩短间接费用的变化，在这基础上计算项目的总费用。

重复以上步骤，直到获得满意的方案为止。

以下通过举例说明。已知项目的每天间接费用为 400 元，利用表 12-6 中的已知资料，按图 12-7 安排进度，项目正常工期为 170 天，对应的项目直接费用为 68 900 元，间接费用为 170×400＝68 000(元)，项目总费用为 136 900 元。这是在正常条件下进行的方案，称为 170 天方案。若要缩短这方案的工期，首先缩短关键路线上直接费用率最小的工作的持续时间，在 170 天方案中关键工作 K,G 的直接费用率是最低。从表中可见这两项工作的持续时间都只能缩短 10 天。由此总工期可以缩短到 170－10－10＝150(天)。按 150 天工期计算，这时总直接费用增加到 68900＋(290×10＋350×10)＝75 300(元)。由于缩短工期，可以减少间接费用 400×20＝8 000(元)，工期为 150 天方案的总费用为 75 300＋60 000＝135 300(元)。与工期 170 天方案相比，可以节省总费用 1 600 元。

但在 150 天方案中已有两条关键路线，即

①→②→④→⑥→⑦→⑧与①→②→⑤→⑦→⑧

如果再缩短工期，工作的直接费用将大幅度增加。例如在 150 天方案的基础上再缩短工期 10 天，成为 140 天方案。这时应选择工作 D，缩短 10 天；工作 H 缩短 5 天(只能缩短 5 天)，工作 E 缩短 5 天。这时直接费用成为 75 300＋400×10＋400×5＋500×5＝83 800(元)。间接费用为 140×400＝56 000(元)，总费用为 139 800 元。显然 140 天方案的总费用比 150 天方案和 170 天方案的总费用都高。综合考虑 150 天方案为最佳方案。计算结果汇总在表 12-7 中。

表 12-6

序号	工作代号	正常持续时间/天	工作直接费用/元	最短工作时间/天	工作直接费用/元	费用率/(元/天)
1	A	60	10 000	60	10 000	/
2	B	45	4 500	30	6 300	120
3	C	10	2 800	5	4 300	300
4	D	20	7 000	10	11 000	400
5	E	40	10 000	35	12 500	500
6	F	18	3 600	10	5 440	230
7	G	30	9 000	20	12 500	350
8	H	15	3 750	10	5 750	400
9	K	25	6 250	15	9 150	290
10	L	35	12 000	35	12 000	/

表 12-7

工期方案	170 天方案	150 天方案	140 天方案
缩短关键工作		K,G	D,H,E
缩短工作持续时间/天		10,10	10,5,5
直接费用/元	68 900	75 300	83 800
间接费用/元	68 000	60 000	56 000
总费用/元	139 600	135 300	139 800

12.5 网络计划软件

12.5.1 概况

20 世纪 60 年代，网络计划图用手工编制时，工作量很大，若遇到有些工作资源配置有变化时，就需要及时修改网络计划图。这是非常麻烦的事，推广应用都存在很多困难。1980 年以后，微机开始普及起来，到 1983 年以后有了在微机上使用的网络计划的软件。这之后经历了几次变革，网络计划的软件随计算机的更新换代而不断改进。目前网络计划技术应用软件的水平，大约处在第五代。大多数软件是 Windows 操作系统。可以处理表格及图形输入方式、图形横道图（Gantt 图）带逻辑连线、单代号网络计划图、完成各种数值计算等。第六代是面向对象的图形智能化的操作，并在不断完善。如我国梦龙科技有限公司开发的智能项目动态控制软件，目前是国内工程领域中用户最多的项目进度控制软件，它极易进行进度计划编制、进度计划优化、进度跟踪反馈、进度分析、控制等各方面工作。

它可同时优化、计划、管理和控制多个项目，可以多方案分析比较、目标计划跟踪、可实时监控。它利用先进的资源平衡来优化资源计划，采用图形化操作、双代号网络图，并同时自动生成 7 种不同模式图、双代号网络图、单代号、逻辑图、单双混合图、横道图、时标网络图，完全根据不同的需要来反映工程各种数据。

将整个项目作为一个系统加以处理，通过网络计划形式对整个系统统筹规划，并进行有效的监控管理，解决项目上最关心的两个问题：进度和费用。

以网络计划为核心技术的梦龙智能项目管理动态控制系统 MrPert，也正好体现了这一点。它很好地解决了项目管理者关心的两个问题：进度和费用，为有效地管理工程提供了极大的方便。

大量的工程项目实践表明，应用网络计划技术管理后可以缩短建设周期 20%，降低工程成本 10%，而编制网络计划所需费用仅为总费用的 0.1%。

在三峡工程、50 周年大阅兵、“神舟”号研制和发射等众多的项目中，梦龙的 MrPert 系统得到很好的应用。

有关网络计划的计算机软件的开发和研究发展很快，现在已有不少商品化的软件。国外的网络计划软件有：Time Line，Project Scheduler，Microsoft Project 等。如 Microsoft Project 有若干不断改进的版本，但基本部分是大同小异的。学习和掌握使用可分阶段进行：初步了解和掌握使用、熟练掌握和灵活使用。它有一套术语，也可以分阶段地熟悉它们。要全部掌握这些软件的使用是需要花一定时间的。编制大型工程项目网络计划图，要掌握网络计划技术，必须要掌握有关网络计划软件的使用。否则网络计划图就失去其实际控制进度的作用。以下作初步介绍，便于读者入门和引起兴趣。

12.5.2 编制网络计划图前的准备工作

1. 项目的工作分解结构(work breakdown structure,WBS)

将一个项目按由粗到细的原则,分解为若干层次,一般是树状结构。分解的粗细程度,取决于使用者的要求。最好将项目的工作分解层次多一些,细一些。这是一项基础工作,要求项目的WBS应当是稳定的,然后确定工作的逻辑关系,列出工作逻辑关系表,如表12-1那样,但是要更详细。一旦经过有关部门审定后,不能随意变动。当将整个项目的WBS输入计算机后,软件可按使用者的需要,随时显示不同层次的纲目和细目。

将工作按逻辑关系表输入计算机后,软件能自动地给WBS的每组成部分指定识别编号。计算机根据工作的识别编号进行运算。并按软件具有的功能,自动完成各种图表的绘制和有关参数的计算。

2. 数据

收集和整理对应项目工作分解结构各层次工作的数据。包括与计算工作持续时间有关的数据,对应各工作的所需要资源及其经济指标,与计算项目经济指标有关的各工作的费用数据,以便进行项目的时间—费用分析。

3. 熟悉有关网络计划软件的术语

不同软件的术语虽然有些不一样,但基本术语是相同的。掌握本章的基本术语后,就比较容易理解不同软件中的术语。另一类是进一步分析用的术语,需要专门学习。

12.5.3 软件的功能

网络计划软件一般都具有:项目清单表、资源清单表、任务清单表、甘特图、网络图、项目分解结构树形图、资源分布图、日历、各种分析报告或报表。图与表之间的数据是动态连接的,保证数据输入或修改的一致性。当在任务清单表中输入项目的各工作与有关数据后,软件自动生成甘特图,网络图根据日历中指定的项目开工日期,按规定的每月或周的工作日和工作班数,计算出各工作的最早开始日期、最迟开始日期、最早结束日期和最迟结束日期,并计算出关键路线和时差。软件具有平衡资源和优化资源利用的功能,根据需要可打印出不同详细程度的网络计划图和有关报表。

1. 界面

一般软件都采用视窗界面,能显示带有时标的甘特图界面、网络计划图界面、资源平衡界面、可指定日期的日历、资源配置表等,显示需要的计算结果,显示关键路线(CPM)。

2. 操作

软件启动后,软件具有操作提示或帮助,便于使用。可以修改初始数据,新增加工作或删改已有工作。各工作的细化或合并,在界面上很容易实现。进行调整软件给出的图表,编制需要的计算关系式和设计输出报表的格式。通过人-机交换,进一步分析网络和优化网络等。

3. 输出结果

打印出网络计划图、甘特图(横道图)、资源平衡图、日历进度计划、分析报表等。有的

软件可允许使用者自行设计报表的格式和打印需要的内容。

习题

12.1　已知下列资料：

工　序	紧前工序	工序时间	工序	紧前工序	工序时间	工序	紧前工序	工序时间
A	*G*,*M*	3	*E*	*C*	5	*I*	*A*,*L*	2
B	*H*	4	*F*	*A*,*E*	5	*K*	*F*,*I*	1
C	—	7	*G*	*B*,*C*	2	*L*	*B*,*C*	7
D	*L*	3	*H*	—	5	*M*	*C*	3

要求：(1) 绘制网络图；

(2) 用图上计算法计算各项时间参数(r 除外)；

(3) 确定关键路线。

12.2　已知下列资料：

工　序	紧前工序	工序时间	工序	紧前工序	工序时间	工序	紧前工序	工序时间
a	—	60	*g*	*b*,*c*	7	*m*	*j*,*k*	5
b	*a*	14	*h*	*e*,*f*	12	*n*	*i*,*l*	15
c	*a*	20	*i*	*f*	60	*o*	*n*	2
d	*a*	30	*j*	*d*,*g*	10	*p*	*m*	7
e	*a*	21	*k*	*h*	25	*q*	*o*,*p*	5
f	*a*	10	*l*	*j*,*k*	10			

要求：(1) 绘制网络图；

(2) 计算各项时间参数；

(3) 确定关键路线。

12.3　已知下列资料：

活　　动	作业时间	紧前活动	正常完成进度的直接费用/百元	赶进度一天所需费用/百元
A	4		20	5
B	8		30	4
C	6	*B*	15	3
D	3	*A*	5	2
E	5	*A*	18	4
F	7	*A*	40	7
G	4	*B*、*D*	10	3
H	3	*E*、*F*、*G*	15	6
合计			153	
工程的间接费用			5(百元/天)	

求出该项工程的最低成本日程。

参考资料

[1] 中国建筑学会建筑统筹管理分会编著.工程网络计划技术规程教程.北京：中国建筑工业出版社,2000

[2] 卢向南.项目计划与控制.北京：机械工业出版社,2004

Part 7 第7篇

排队论

排队是在日常生活中经常遇到的现象，如顾客到商店购买物品、病人到医院看病常常要排队。此时要求服务的人数超过服务机构（服务台、服务员等）的容量，也就是说，到达的顾客不能立即得到服务，因而出现了排队现象。这种现象不仅在个人日常生活中出现，而且很多服务系统，如电话服务热线的占线问题，车站、码头等交通枢纽的车船堵塞和疏导，故障机器的停机待修，水库水量的存储调节等都存在有形或无形的排队现象。以下将要求服务的对象统称为顾客。由于顾客到达和服务时间的随机性，可以说排队现象几乎是不可避免的。

如果增添服务设备，就要增加投资或发生空闲浪费；如果服务设备太少，排队现象就会严重，对顾客个人和对社会都会带来不利影响。因此，管理人员必须考虑如何在这两者之间取得平衡，经常检查目前处理是否得当，研究今后改进对策，以期提高服务质量，降低成本。

排队论(queuing theory)也称**随机服务系统理论**，就是为解决上述问题而发展起来的一门学科，它研究的内容有下列三部分。

(1) 性态问题，即研究各种排队系统的概率规律性，主要是研究队长分布、等待时间分布和忙期分布等，包括了瞬态和稳态两种情形。

(2) 最优化问题，又分静态最优和动态最优，前者指最优设计，后者指现有排队系统的最优运营。

(3) 排队系统的统计推断，即判断一个给定的排队系统符合于哪种模型，以便根据排队理论进行分析研究。

这里将介绍排队论的一些基本知识，分析几种常见的排队模型，最后将介绍排队系统的最优化问题。

CHAPTER 13

第13章

排　队　论

13.1　基本概念

13.1.1　排队过程的一般表示

图 13-1 就是排队过程的一般模型。各个顾客由顾客源(总体)出发,到达服务机构(服务台、服务员)前排队等候接受服务,服务完成后就离开。排队结构指队列的数目和排列方式,排队规则和服务规则是说明顾客在排队系统中按怎样的规则、次序接受服务的。我们所说的排队系统就指图中虚线所包括的部分。

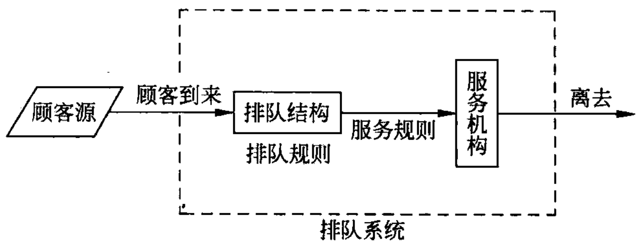


图　13-1

在现实中的排队现象是多种多样的,对上面所说的"顾客"和"服务员",要作广泛的了解。它可以是人,也可以是非生物;队列可以是具体地排列,也可以是无形的(例如向电话交换台要求通话的呼唤);顾客可以走向服务机构,也可以相反(如送货上门)。下面举一些例子说明现实中形形色色的排队系统(见表 13-1)。

表　13-1

到达的顾客	要求服务的内容	服 务 机 构
不能运转的机器	修理	修理技工
修理技工	领取修配零件	发放修配零件的管理员
发放修配零件的管理员	购进修配零件	采购员
病人	诊断或动手术	医生(或包括手术台)
电话呼唤	通话	交换台
文件稿	打字	打字员

续表

到达的顾客	要求服务的内容	服 务 机 构
提货单	提取存货	仓库管理员
到达机场上空的飞机	降落	跑道
驶入港口的货船	装(卸)货	装(卸)货码头(泊位)
上游河水进入水库	放水,调整水位	水闸管理员
进入我方阵地的敌机	我方高射炮进行射击	我方高射炮
下火车后的出租车乘客	乘坐出租车	火车站出租车管理员

13.1.2 排队系统的组成和特征

一般的排队系统都有三个基本组成部分:①输入过程;②排队规则;③服务机构。现在分别说明各部分的特征。

1. 输入过程

输入即指顾客到达排队系统,可能有下列各种不同情况,当然这些情况并不是彼此排斥的。

(1) 顾客的总体(称为顾客源)的组成可能是有限的,也可能是无限的。上游河水流入水库可以认为总体是无限的,工厂内停机待修的机器显然是有限的总体。

(2) 顾客到来的方式可能是一个一个的,也可能是成批的。例如到餐厅就餐就有单个到来的顾客和受邀请来参加宴会的成批顾客,我们将只研究单个到来的情形。

(3) 顾客相继到达的间隔时间可以是确定型的,也可以是随机型的。如在自动装配线上装配的各部件就必须按确定的时间间隔到达装配点,定期运行的班车、班轮、班机的到达也都是确定型的。但一般到商店购物的顾客、到医院诊病的病人、通过路口的车辆等,它们的到达都是随机型的。对于随机型的情形,要知道单位时间内的顾客到达数或相继到达的间隔时间的概率分布(图 13-2)。

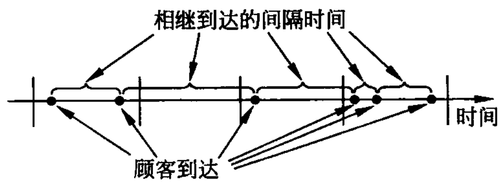


图 13-2

(4) 顾客的到达可以是**相互独立的**,就是说,以前的到达情况对以后顾客的到来没有影响,否则就是有关联的。例如,工厂内的机器在一个短的时间区间内出现停机(顾客到达)的概率就受已经待修或被修理的机器数目的影响。我们主要讨论的是相互独立的情形。

(5) 输入过程可以是**平稳**的，或称对时间是齐次的，是指描述相继到达的间隔时间分布和所含参数(如期望值、方差等)都是与时间无关的，否则称为非平稳的。我们主要讨论平稳的情形，非平稳情形的数学处理是很困难的。

2. 排队规则

(1) 顾客到达时，如所有服务台都正被占用，在这种情形下顾客可以随即离去，也可以排队等候。随即离去的称为**即时制**或称**损失制**，因为这将失掉许多顾客；排队等候的称为**等待制**。如一位顾客到达某一餐馆，如果全部座位都已满，他即离开该餐馆往别处吃饭属于前者，如果他选择等待则属于后者。

对于等待制，为顾客进行服务的次序可以采用下列各种规则：先到先服务，后到先服务，随机服务，有优先权的服务等。

先到先服务，即按到达次序接受服务，这是最通常的情形。

后到先服务，如乘用电梯的顾客常是后入先出的。仓库中存放的厚钢板也是如此。在情报系统中，最后到达的信息往往是最有价值的，因而常采用后到先服务(指被采用)的规则。

随机服务，指服务员从等待的顾客中随机地选取其一进行服务，而不管到达的先后，如乘客(视为服务台)在停车场随机选乘(乘车视为服务)一辆出租车(视为等候服务的顾客)。

有优先权的服务，如医院对于病情严重的患者将给予优先治疗。

(2) 从占有的空间来看，队列可以排在具体的处所(如售票处、候诊室等)，也可以是抽象的(如向电话交换台要求通话的呼唤)。由于空间的限制或其他原因，有的系统要规定容量(即允许进入排队系统的顾客数)的最大限量；有的没有这种限制(即认为容量可以是无限的)。

(3) 从队列的数目看，可以是**单列**，也可以是**多列**。在多列的情形，各列间的顾客有的可以互相转移，有的不能(如用绳子或栏杆隔开)。有的排队顾客因等候时间过长而中途退出，有的不能退出(如高速公路上的汽车流)，必须坚持到被服务为止。我们将只讨论各队列间不能互相转移，也不能中途退出的情形。

3. 服务机构

从机构形式和工作情况来看有以下几种情况。

(1) 服务机构可以没有服务员，也可以有一个或多个服务员(服务台、通道、窗口等)。例如，在敞架售书的书店，顾客选书时就没有服务员，但交款时可能有多个服务员。

(2) 在有多个服务台的情形中，它们可以是平行排列(并列)的，可以是前后排列(串列)的，也可以是混合的。图 13-3 说明了这些情形。

图 13-3 中(a)是单队—单服务台的情形；(b)是多队—多服务台(并列)的情形；(c)是单队—多服务台(并列)的情形；(d)是多服务台(串列)的情形；(e)是多服务台(混合)的情形。

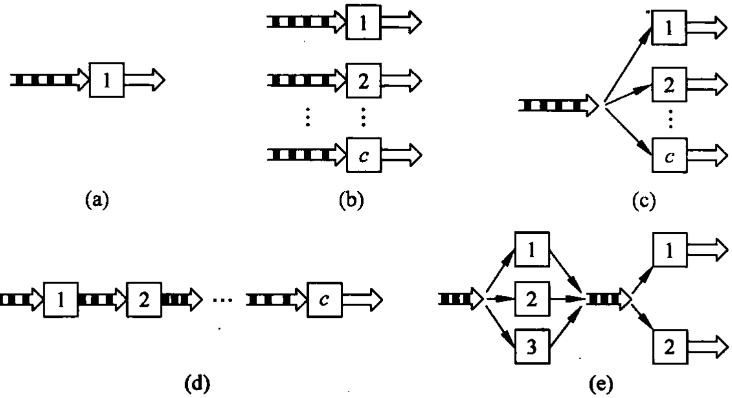


图 13-3

(3) 服务方式可以对单个顾客进行,也可以对成批顾客进行,公共汽车对在站台等候的顾客就成批进行服务。我们将只研究对单个顾客的服务方式。

(4) 和输入过程一样,服务时间也分**确定型**的和**随机型**的。自动冲洗汽车的装置对每辆汽车冲洗(服务)的时间就是确定型的,但大多数情形的服务时间是随机型的。对于随机型的服务时间,需要知道它的概率分布。

如果输入过程,即相继到达的间隔时间和服务时间二者都是确定型的,那么问题就太简单了。因此,在排队论中所讨论的是二者至少有一个是随机型的情形。

(5) 和输入过程一样,服务时间的分布我们总假定是平稳的,即分布的期望值、方差等参数都不受时间的影响。

13.1.3 排队模型的分类

D. G. Kendall 在 1953 年提出排队模型分类方法,对分类方法影响最大的特征有三个,即

(1) 相继顾客到达间隔时间的分布;

(2) 服务时间的分布;

(3) 服务台的个数。

按照这三个特征分类,并用一定符号表示,称为 Kendall 记号。这只对并列的服务台(如果服务台是多于一个的话)的情形,他用的符号形式是

$$X/Y/Z$$

其中,X 处填写表示相继到达间隔时间的分布;

Y 处填写表示服务时间的分布;

Z 处填写并列的服务台的数目。

表示相继到达间隔时间和服务时间的各种分布的符号是:

M——负指数分布(M 是 Markov 的字头,因为负指数分布具有无记忆性,即 Markov 性);

D——确定型(deterministic);

E_k——k 阶爱尔朗(erlang)分布;

GI—— 一般相互独立(general independent)的时间间隔的分布;

G—— 一般(general)服务时间的分布。

例如,$M/M/1$ 表示相继到达间隔时间为负指数分布、服务时间为负指数分布、单服务台的模型;$D/M/c$ 表示确定的到达间隔、服务时间为负指数分布、c 个平行服务台(但顾客是一队)的模型。

后来,在 1971 年一次关于排队论符号标准化会议上决定,将 Kendall 符号扩充成为:$X/Y/Z/A/B/C$ 形式,其中前三项意义不变,而后三项意义分别是:

A 处填写系统容量限制 N;

B 处填写顾客源数目 m;

C 处填写服务规则,如先到先服务(FCFS),后到先服务(LCFS)等。

并约定,如略去后三项,即指 $X/Y/Z/\infty/\infty/$FCFS 的情形。在本书中,因只讨论先到先服务 FCFS 的情形,所以略去第六项。

13.1.4 排队问题的求解

一个实际问题作为排队问题求解时,首先要研究它属于哪个模型,其中只有**顾客到达的间隔时间分布**和**服务时间的分布**需要实测的数据来确定,其他因素都是在问题提出时给定的。

解排队问题的目的,是研究排队系统运行的效率,估计服务质量,确定系统参数的最优值,以决定系统结构是否合理、研究设计改进措施等。所以必须确定用以判断系统运行优劣的基本数量指标,解排队问题就是首先求出这些数量指标的概率分布或特征数。这些指标通常是:

(1) **队长**,指在系统中的顾客数,它的期望值记作 L_s;

排队长(队列长),指在系统中排队等待服务的顾客数,它的期望值记作 L_q;

$$\begin{bmatrix}\text{系统中}\\\text{顾客数}\end{bmatrix}=\begin{bmatrix}\text{在队列中等待}\\\text{服务的顾客数}\end{bmatrix}+\begin{bmatrix}\text{正被服务}\\\text{的顾客数}\end{bmatrix}$$

一般情形,L_s(或 L_q)越大,说明服务率越低,排队成龙,是顾客最厌烦的。

(2) **逗留时间**,指一个顾客在系统中的停留时间,它的期望值记作 W_s;

等待时间,指一个顾客在系统中排队等待的时间,它的期望值记作 W_q,

$$[\text{逗留时间}]=[\text{等待时间}]+[\text{服务时间}]$$

在机器故障问题中,无论是等待修理或正在修理都使工厂受到停工的损失。所以逗留时间(停工时间)是主要的。但一般购物、诊病等问题中仅仅等待时间常是顾客们所关心的。

此外，还有**忙期**(busy period)指从顾客到达空闲服务机构起到服务机构再次为空闲止这段时间长度，即服务机构连续繁忙的时间长度，它关系到服务员的工作强度。忙期和一个忙期中平均完成服务顾客数都是衡量服务机构效率的指标。

在即时制或排队有限制的情形，还有由于顾客被拒绝而使企业受到损失的**损失率**以及以后经常遇到的**服务强度**等，这些都是很重要的指标。

计算这些指标的基础是表达系统状态的概率。所谓**系统的状态**即指系统中的顾客数(其期望值即 L_s)，如果系统中有 n 个顾客就称系统的状态是 n，它的可能值是：

(1) 队长没有限制时，$n=0,1,2,\cdots$

(2) 队长有限制，最大数为 N 时，$n=0,1,2,\cdots,N$

(3) 即时制，服务台个数是 c 时，$n=0,1,2,\cdots,c$

后者，状态 n 又表示正在工作(繁忙)的服务台数。

这些状态的概率一般是随时刻 t 而变化，所以在时刻 t、系统状态为 n 的概率用 $P_n(t)$ 表示。

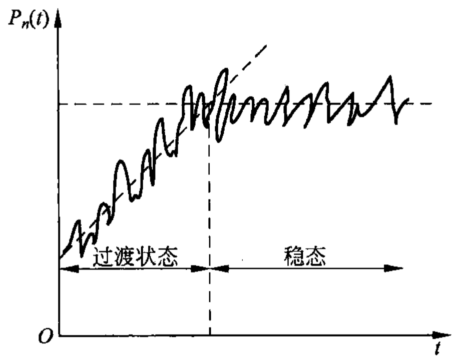


图 13-4

求状态概率 $P_n(t)$ 的方法，首先要建立含 $P_n(t)$ 的关系式见图 13-4，因为 t 是连续变量，而 n 只取非负整数，所以建立的 $P_n(t)$ 的关系式一般是微分差分方程(关于 t 的微分方程，关于 n 的差分方程)。方程的解称为**瞬态**(或称**过渡状态**)(transient state)解。求瞬态解是不容易的，一般地，即使求出也很难利用，因此我们常考虑它的极限(如果存在的话)

$$\lim_{t\to\infty} P_n(t)=P_n$$

称为稳态(steady state)，或称**统计平衡状态**(statistical equilibrium state)的解。

稳态的物理含义是，当系统运行了无限长的时间之后，初始($t=0$)出发状态的概率分布($P_n(0),n\geqslant0$)的影响将消失，而且系统的状态概率分布不再随时间变化。当然，在实际应用中大多数问题系统会很快趋于稳态，而无须等到 $t\to\infty$ 以后。但永远达不到稳态的情形也确实存在的。

求稳态概率 P_n 时，并不一定求 $t\to\infty$ 时 $P_n(t)$ 的极限，而只需令导数 $P_n'(t)=0$ 即可。我们以下着重研究稳态的情形。

13.2 到达间隔的分布和服务时间的分布

解决排队问题首先要根据原始资料作出顾客到达间隔和服务时间的经验分布，然后按照统计学的方法(例如 χ^2 检验法)以确定合于哪种理论分布，并估计它的参数值。本节

先举例说明经验分布，然后介绍常见的理论分布——泊松分布、负指数分布和爱尔朗(Erlang)分布。

13.2.1 经验分布

现在举例说明原始资料的整理。

例 1 大连某港区 2011 年载货 500 吨以上船舶到达（不包括定期到达的船舶）逐日记录见表 13-2。

表 13-2

月＼日	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	8	10	4	1	0	1	4	4	7	0	3	2	4	0	2	2
2	2	5	5	2	7	3	6	2	3	2	2	2	2	3	4	6
3	3	3	5	2	4	7	4	3	3	4	3	2	4	5	3	2
4	6	2	1	5	4	3	4	3	7	5	5	1	1	3	7	4
5	7	4	1	2	2	3	4	3	5	3	2	7	5	4	3	7
6	1	4	2	6	5	2	7	4	3	3	1	3	4	5	5	3
7	4	1	1	3	3	4	8	4	4	4	3	5	3	0	7	4
8	4	3	4	4	3	1	2	5	5	3	5	4	3	6	1	3
9	3	4	4	1	7	7	0	7	2	3	0	6	6	2	2	3
10	5	2	1	1	6	6	4	6	2	3	6	7	3	5	3	2
11	5	4	3	5	2	3	4	4	3	1	3	3	3	4	5	0
12	6	5	5	2	1	5	4	2	2	2	3	1	7	5	1	4

月＼日	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	5	4	2	1	1	3	6	2	3	4	4	2	0	3	1
2	2	2	6	5	2	1	5	3	5	5	2	2			
3	1	2	3	8	2	3	3	2	5	4	7	2	4	1	3
4	5	4	4	4	6	5	1	4	4	1	4	3	6	4	
5	2	1	6	3	2	5	2	5	4	2	4	4	4	4	2
6	5	2	4	3	3	3	6	3	5	5	6	2	1	4	
7	1	3	4	4	2	3	5	5	1	2	4	3	4	6	3
8	2	0	6	3	4	6	4	4	5	1	2	8	4	5	1
9	4	4	5	3	1	4	6	1	2	3	5	0	2	4	
10	6	2	5	1	0	7	9	3	2	5	1	7	3	5	3
11	2	1	0	3	6	6	3	5	5	1	2	2	2	4	
12	6	7	3	1	2	1	3	4	1	3	9	3	3	1	4

摘自《港口装卸》，1980 第 5 期，第 5 页。

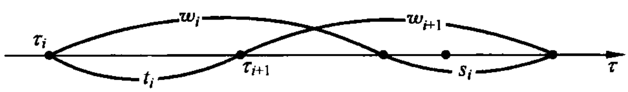
将表 13-2 整理成船舶到达数的分布表(表 13-3)。

表 13-3 船舶到达数分布表(大连港大港区 1979)

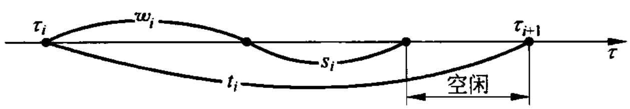
船舶到达数 n	频　数	频　率/%
0	12	0.033
1	43	0.118
2	64	0.175
3	74	0.203
4	71	0.195
5	49	0.134
6	26	0.071
7	19	0.052
8	4	0.011
9	2	0.005
10 以上	1	0.003
合　计	365	1.000

$$\text{平均到达率}=\frac{\text{到达总数}}{\text{总天数}}=\frac{1\,271}{365}=3.48(\text{艘/天})$$

更原始的资料是记录各顾客到达的时刻和对各顾客的服务时间。以 τ_i 表示第 i 号顾客到达的时刻，以 s_i 表示对它的服务时间，这样可算出相继到达的间隔时间 $t_i(t_i=\tau_{i+1}-\tau_i)$ 和排队等待时间 w_i，它们的关系见图 13-5。



(a) $w_i+s_i-t_i>0$



(b) $w_i+s_i-t_i<0$

图 13-5

从图 13-5 中看出

间隔　　$t_i=\tau_{i+1}-\tau_i$

等待时间　　$$w_{i+1}=\begin{cases} w_i+s_i-t_i, & \text{当 } w_i+s_i-t_i>0 \\ 0, & \text{当 } w_i+s_i-t_i<0 \end{cases} \qquad (13\text{-}1)$$

例 2 某服务机构是单服务台，先到先服务，对 41 个顾客记录到达时刻 τ 和服务时间 s(单位为分钟)如表 13-4，在表中以第 1 号顾客到达时刻为 0。全部服务时间为 127 分钟。

表 13-4

(1)i	(2)τ_i	(3)s_i	(4)t_i	(5)w_i	(1)i	(2)τ_i	(3)s_i	(4)t_i	(5)w_i
1	0	5	2	0	22	83	3	3	2
2	2	7	4	3	23	86	6	2	2
3	6	1	5	6	24	88	5	4	6
4	11	9	1	2	25	92	1	3	7
5	12	2	7	10	26	95	3	6	5
6	19	4	3	5	27	101	2	4	2
7	22	3	4	6	28	105	2	1	0
8	26	3	10	5	29	106	1	3	1
9	36	1	2	0	30	109	2	5	0
10	38	2	7	0	31	114	1	2	0
11	45	5	2	0	32	116	8	1	0
12	47	4	2	3	33	117	4	4	7
13	49	1	3	5	34	121	2	6	7
14	52	2	9	3	35	127	1	2	3
15	61	1	1	0	36	129	6	1	2
16	62	2	3	0	37	130	3	3	7
17	65	1	5	0	38	133	5	2	7
18	70	3	2	0	39	135	2	4	10
19	72	4	8	1	40	139	4	3	8
20	80	3	1	0	41	142	1		9
21	81	2	2	2					

各栏意义：

(1)顾客编号 i；(2)到达时刻 τ_i；(3)服务时间 s_i；以上三栏是原始记录；(4)到达间隔 t_i；(5)排队等待时间 w_i，这两栏是通过式(13-1)计算得到的。

现将上面的原始记录整理成到达间隔分布表(表 13-5)和服务时间分布表(表 13-6)。

表 13-5

到达间隔/分钟	次　　数
1	6
2	10
3	8
4	6
5	3
6	2
7	2
8	1
9	1
10 以上	1
合　计	40

表 13-6

服务时间/分钟	次　　数
1	10
2	10
3	7
4	5
5	4
6	2
7	1
8	1
9 以上	1
合　计	41

平均间隔时间＝142/40＝3.55(分钟/人)

平均到达率＝41/142＝0.28(人/分钟)

平均服务时间＝127/41＝3.12(分钟/人)

平均服务率＝41/127＝0.32(人/分钟)

下面介绍经常用的几个理论分布。

13.2.2 泊松流

设 $N(t)$ 表示在时间区间 $[0,t)$ 内到达的顾客数($t>0$)

令 $P_n(t_1,t_2)$ 表示在时间区间 $[t_1,t_2)$($t_2>t_1$)内有($n\geqslant 0$)个顾客到达(这当然是随机事件)的概率，即

$$P_n(t_1,t_2)=P\{N(t_2)-N(t_1)=n\} \quad (t_2>t_1,n\geqslant 0)$$

当 $P_n(t_1,t_2)$ 合于下列三个条件时，称顾客的到达形成**泊松流**。这三个条件是：

(1) 在不相重叠的时间区间内顾客到达数是相互独立的，我们称这性质为无后效性。

(2) 对充分小的 Δt，在时间区间 $[t,t+\Delta t)$ 内有 1 个顾客到达的概率与 t 无关，而约与区间长 Δt 成正比，即

$$P_1(t,t+\Delta t)=\lambda\Delta t+o(\Delta t) \tag{13-2}$$

其中，$o(\Delta t)$，当 $\Delta t\to 0$ 时，是关于 Δt 的高阶无穷小。$\lambda>0$ 是常数，它表示单位时间有一个顾客到达的概率，称为概率强度。

(3) 对于充分小的 Δt，在时间区间 $[t,t+\Delta t)$ 内有 2 个或 2 个以上顾客到达的概率极小，以至于可以忽略，即

$$\sum_{n=2}^{\infty} P_n(t,t+\Delta t) = o(\Delta t) \tag{13-3}$$

在上述条件下，我们研究顾客到达数 n 的概率分布。

由条件(2)，我们总可以取时间由 0 算起，并简记 $P_n(0,t)=P_n(t)$，$P_n(t)$ 表示长为 t 的时间区间内到达 n 个顾客的概率。

由条件(2)和(3)，容易推得在 $[t,t+\Delta t)$ 区间内没有顾客到达的概率

$$P_0(t,t+\Delta t)=1-\lambda\Delta t+o(\Delta t) \tag{13-4}$$

在求 $P_n(t)$ 时，用通常建立未知函数的微分方程的方法，先求未知函数 $P_n(t)$ 由时刻 t 到 $t+\Delta t$ 的改变量，从而建立 t 时刻的概率分布与 $t+\Delta t$ 时刻概率分布的关系方程。

对于区间 $[0,t+\Delta t)$，可分成两个互不重叠的区间 $[0,t)$ 和 $[t,t+\Delta t)$。现在到达总数是 n，分别出现在这两区间上，不外下列三种情况。各种情况出现个数和概率见表 13-7。

在 $[0,t+\Delta t)$ 内到达 n 个顾客应是表中三种互不相容的情况之一，所以概率 $P_n(t+\Delta t)$ 应是表中三个概率之和(各 $o(\Delta t)$ 合为一项)

$$P_n(t+\Delta t)=P_n(t)(1-\lambda\Delta t)+P_{n-1}(t)\lambda\Delta t+o(\Delta t)$$

$$\frac{P_n(t+\Delta t)-P_n(t)}{\Delta t}=-\lambda P_n(t)+\lambda P_{n-1}(t)+\frac{o(\Delta t)}{\Delta t}$$

表 13-7

情况	区间 [0, t) 个数	[0, t) 概率	[t, t+Δt) 个数	[t, t+Δt) 概率	[0, t+Δt) 个数	[0, t+Δt) 概率
(A)	n	$P_n(t)$	0	$1-\lambda\Delta t+o(\Delta t)$	n	$P_n(t)(1-\lambda\Delta t+o(\Delta t))$
(B)	$n-1$	$P_{n-1}(t)$	1	$\lambda\Delta t$	n	$P_{n-1}(t)\lambda\Delta t$
(C)	$n-2$	$P_{n-2}(t)$	2		n	
	$n-3$	$P_{n-3}(t)$	3	$o(\Delta t)$	n	$o(\Delta t)$
	⋮	⋮	⋮		⋮	
	0	$P_0(t)$	n		n	

令 $\Delta t\to 0$，得下列方程，并注意到初始条件，则有

$$\begin{cases}\dfrac{\mathrm{d}P_n(t)}{\mathrm{d}t}=-\lambda P_n(t)+\lambda P_{n-1}(t), \quad n\geqslant 1\\ P_n(o)=0;\end{cases} \tag{13-5}$$

当 $n=0$ 时，没有 (B)，(C) 两种情况，所以得

$$\begin{cases}\dfrac{\mathrm{d}P_0(t)}{\mathrm{d}t}=-\lambda P_0(t)\\ P_0(o)=1\end{cases} \tag{13-6}$$

解式(13-5)和式(13-6)，就得

$$\boxed{\begin{aligned}&P_n(t)=\frac{(\lambda t)^n}{n!}e^{-\lambda t},\\ &t>0 \qquad n=0,1,2,\cdots\end{aligned}}\text{①} \tag{13-7}$$

---

① 证明时先解式(13-6)得 $P_0(t)=e^{-\lambda t}$，
然后在方程式(13-5)两边乘积分因子 $e^{\lambda t}$，并迁项

$$e^{\lambda t}\frac{\mathrm{d}P_n(t)}{\mathrm{d}t}+\lambda P_n(t)e^{\lambda t}=\lambda e^{\lambda t}P_{n-1}(t)$$

$$\frac{\mathrm{d}}{\mathrm{d}t}[P_n(t)e^{\lambda t}]=\lambda P_{n-1}(t)e^{\lambda t}$$

积分得

$$P_n(t)e^{\lambda t}=\lambda\int_u^t P_{n-1}(t_1)e^{\lambda t_1}\,\mathrm{d}t_1$$

依次代入 $n=1,2,\cdots$

$$n=1,\quad P_1(t)e^{\lambda t}=\lambda\int_0^t e^{-\lambda t_1}e^{\lambda t_1}\,\mathrm{d}t_1=\lambda t,\quad P_1(t)=\lambda te^{-\lambda t}$$

$$n=2,\qquad P_2(t)e^{\lambda t}=\lambda\int_0^t \lambda t_1e^{-\lambda t_1}e^{\lambda t_1}\,\mathrm{d}t_1=\frac{1}{2}\lambda^2t^2$$

所以

$$P_2(t)=\frac{\lambda^2t^2}{2!}e^{-\lambda t}$$

递推，即可得式(13-7)。 证毕。

式(13-7)就是在概率论中的随机变量 $N(t)$服从泊松分布的定义,只是一般取 $t=1$。$N(t)$的数学期望和方差分别是

$$E[N(t)]=\lambda t;\quad \mathrm{Var}[N(t)]=\lambda t \tag{13-8}$$

期望值和方差相等,是泊松分布的一个重要特征,我们可以利用它对一个经验分布是否合于泊松分布进行初步的识别。

13.2.3 负指数分布

随机变量 T 的概率密度若是

$$f_T(t)=\begin{cases}\lambda e^{-\lambda t}, & t\geqslant 0\\ 0, & t<0\end{cases} \tag{13-9}$$

则称 T 服从负指数分布。它的分布函数是

$$F_T(t)=\begin{cases}1-e^{-\lambda t}, & t\geqslant 0\\ 0, & t<0\end{cases} \tag{13-10}$$

数学期望 $E[T]=\dfrac{1}{\lambda}$;方差 $\mathrm{Var}[T]=\dfrac{1}{\lambda^2}$;标准差 $\sigma[T]=\dfrac{1}{\lambda}$

负指数分布有下列性质:

(1) 由条件概率公式容易证明

$$P\{T>t+s\mid T>s\}=P\{T>t\} \tag{13-11}$$

这个性质称为无记忆性或马尔可夫性。若 T 表示排队系统中顾客到达的间隔时间,那么这个性质说明现在考虑一个顾客到来还需的时间 t 与之前已过去的时间 s 无关,所以说这一情形下的顾客到达是纯随机的。

(2) 当输入过程是泊松流时,那么顾客相继到达的间隔时间 T 必须服从负指数分布。这是因为对于泊松流,在$[0,t)$区间内至少有 1 个顾客到达的概率是

$$1-P_0(t)=1-e^{-\lambda t},\ t>0$$

而这概率又可表示为

$$P\{T\leqslant t\}=F_T(t)$$

结合式(13-10),这性质得到证明。

因此,相继到达的间隔时间是独立且相同的负指数分布(密度函数为 $\lambda e^{-\lambda t},t\geqslant 0$),与输入过程为泊松流(参数为 λ)是等价的。所以在 Kendall 记号中就都用 M 表示。

对于泊松流,λ 表示单位时间平均到达的顾客数,所以 $1/\lambda$ 就表示相继顾客到达平均间隔时间,而这正和 $E[T]$的意义相符。

服务时间 v 的分布即对一顾客的服务时间,也就是在忙期相继离开系统的两顾客的间隔时间,有时也服从负指数分布。这时设它的分布函数和密度分别是

$$F_v(t)=1-e^{-\mu t},\quad f_v(t)=\mu e^{-\mu t} \tag{13-12}$$

其中，μ 表示单位时间能被服务完成的顾客数，称为平均服务率，而 $\frac{1}{\mu}=E(v)$ 表示一个顾客的平均服务时间，这里平均就是期望值。

13.2.4 爱尔朗分布

设 $v_1, v_2, \cdots, v_k$ 是 k 个相互独立的随机变量，服从相同参数 $k\mu$ 的负指数分布，那么

$$T=v_1+v_2+\cdots+v_k$$

的概率密度是

$$b_k(t)=\frac{\mu k(\mu k t)^{k-1}}{(k-1)!}e^{-\mu k t}, t>0 \tag{13-13}$$

（证明略）我们说 T 服从 k 阶爱尔朗分布。

$$E[T]=\frac{1}{\mu};\ \mathrm{Var}[T]=\frac{1}{k\mu^2} \tag{13-14}$$

这是因为

$$E[v_i]=\frac{1}{k\mu}, i=1,2,\cdots,k$$

所以

$$E[T]=\sum_{i=1}^{k}E(v_i)=\frac{1}{\mu}$$

例如串列的 k 个服务台，每台服务时间相互独立，服从相同的负指数分布（参数 $k\mu$），那么一顾客走完这 k 个服务台总共所需要服务时间就服从上述的 k 阶爱尔朗分布。

注意 爱尔朗分布族提供更为广泛的模型类，比指数分布有更大的适应性。事实上，当 $k=1$ 时，爱尔朗分布化为负指数分布，这可看成是完全随机的；当 k 增大时，爱尔朗分布的图形逐渐变为对称的；当 $k\geqslant 30$ 时爱尔朗分布近似于正态分布；$k\to\infty$时，由式(13-14)看出 $\mathrm{Var}[T]\to 0$，因此这时爱尔朗分布化为确定型分布（见图 13-6），所以一般 k 阶爱尔朗分布可看成完全随机与完全确定的中间型，能对现实世界提供更为广泛的适应性。

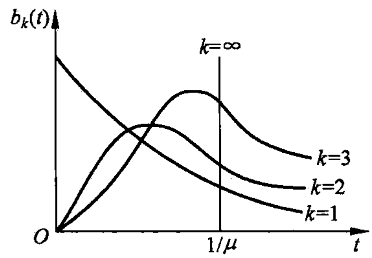


图 13-6

13.3 单服务台负指数分布排队系统的分析

在本节中讨论单服务台的排队系统，它的输入过程服从泊松分布过程，服务时间服从负指数分布。按以下三种情形讨论。

(1) 标准的 $M/M/1$ 模型，即$(M/M/1/\infty/\infty)$；

(2) 系统的容量有限制，即$(M/M/1/N/\infty)$；

(3) 顾客源为有限,即($M/M/1/\infty/m$)。

13.3.1 标准的 $M/M/1$ 模型($M/M/1/\infty/\infty$)

标准的 $M/M/1$ 模型是指适合下列条件的排队系统:

(1) 输入过程——顾客源是无限的,顾客单个到来,相互独立,一定时间的到达数服从泊松分布,到达过程也是平稳的。

(2) 排队规则——单队,且对队长没有限制,先到先服务。

(3) 服务机构——单服务台,各顾客的服务时间是相互独立的,服从相同的负指数分布。

此外,还假定到达间隔时间和服务时间是相互独立的。

在分析标准的 $M/M/1$ 模型时,首先要求出系统在任意时刻 t 的状态为 n(系统中有 n 个顾客)的概率 $P_n(t)$,它决定了系统运行的特征。

因已知到达规律服从参数为 λ 的泊松过程,服务时间服从参数为 μ 的负指数分布,所以在$[t,t+\Delta t)$时间区间内分为:

(1) 有 1 个顾客到达的概率为 $\lambda\Delta t+o(\Delta t)$;没有顾客到达的概率就是 $1-\lambda\Delta t+o(\Delta t)$。

(2) 当有顾客在接受服务时,1 个顾客被服务完了(离去)的概率是 $\mu\Delta t+o(\Delta t)$,没有离去的概率就是 $1-\mu\Delta t+o(\Delta t)$。

(3) 多于一个顾客的到达或离去的概率是 $o(\Delta t)$,是可以忽略的。

在时刻 $t+\Delta t$,系统中有 n 个顾客($n>0$)存在表 13-8 的四种情况(到达或离去是 2 个以上的没列入):

表 13-8

情　况	在时刻 t 顾客数	在区间$[t,\ t+\Delta t)$		在时刻 $t+\Delta t$ 顾客数
		到　达	离　去	
(A)	n	×	×	n
(B)	$n+1$	×	○	n
(C)	$n-1$	○	×	n
(D)	n	○	○	n

○表示发生(1 个);×表示没有发生。

它们的概率分别是(略去 $o(\Delta t)$):

情况(A)　　$P_n(t)(1-\lambda\Delta t)(1-\mu\Delta t)$

情况(B)　　$P_{n+1}(t)(1-\lambda\Delta t)\mu\Delta t$

情况(C)　　$P_{n-1}(t)\cdot\lambda\Delta t(1-\mu\Delta t)$

情况(D)　　$P_n(t)\cdot\lambda\Delta t\cdot\mu\Delta t$

由于这四种情况是互不相容的,所以 $P_n(t+\Delta t)$应是这四项之和,即(将关于 Δt 的高阶无穷小合成一项)

$$P_n(t+\Delta t)=P_n(t)(1-\lambda\Delta t-\mu\Delta t)+P_{n+1}(t)\mu\Delta t+P_{n-1}(t)\lambda\Delta t+o(\Delta t)$$

$$\frac{P_n(t+\Delta t)-P_n(t)}{\Delta t}=\lambda P_{n-1}(t)+\mu P_{n+1}(t)-(\lambda+\mu)P_n(t)+\frac{o(\Delta t)}{\Delta t}$$

令 $\Delta t\to 0$,得关于 $P_n(t)$的微分差分方程

$$\frac{\mathrm{d}P_n(t)}{\mathrm{d}t}=\lambda P_{n-1}(t)+\mu P_{n+1}(t)-(\lambda+\mu)P_n(t)\quad n=1,2,\cdots \tag{13-15}$$

当 $n=0$,则只有表 13-8 中(A),(B)两种情况,即

$$P_0(t+\Delta t)=P_0(t)(1-\lambda\Delta t)+P_1(t)(1-\lambda\Delta t)\mu\Delta t$$

同理求得

$$\frac{\mathrm{d}P_0(t)}{\mathrm{d}t}=-\lambda P_0(t)+\mu P_1(t) \tag{13-16}$$

这样系统状态(n)随时间变化的过程是称为生灭过程的一个特殊情形。

解方程式(13-15)和式(13-16)是很麻烦的,求得的解(瞬态解)中因为含有修正的贝塞耳函数,也不便于应用,我们只研究稳态的情况,这时 $P_n(t)$与 t 无关,可写成 P_n,它的导数为 0。由式(13-15)和式(13-16)可得

$$\begin{cases}-\lambda P_0+\mu P_1=0 & (13\text{-}17)\\ \lambda P_{n-1}+\mu P_{n+1}-(\lambda+\mu)P_n=0\quad n\geqslant 1 & (13\text{-}18)\end{cases}$$

这是关于 P_n 的差分方程。它表明了各状态间的转移关系,用图 13-7 表示。

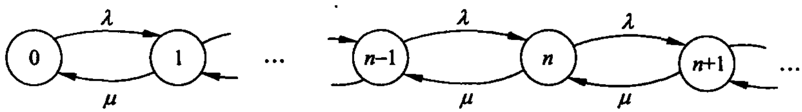


图　13-7

由图 13-7 可见,状态 0 转移到状态 1 的转移率为 λP_0,状态 1 转移到状态 0 的转移率为 μP_1。对状态 0 必须满足以下平衡方程

$$\lambda P_0=\mu P_1$$

同样对任何 $n\geqslant 1$ 的状态,可得到式(13-18)的平衡方程。求解式(13-17)得

$$P_1=(\lambda/\mu)P_0$$

将它代入式(13-18),令 $n=1$,

$$\mu P_2=(\lambda+\mu)(\lambda/\mu)P_0-\lambda P_0;\qquad 所以\quad P_2=(\lambda/\mu)^2P_0$$

同理依次推得

$$P_n=(\lambda/\mu)^n P_0$$

今设 $\rho=\dfrac{\lambda}{\mu}<1$(否则队列将排至无限远),又由概率的性质知

$$\sum_{n=0}^{\infty} P_n = 1$$

将 P_n 的关系代入

$$P_0\sum_{n=0}^{\infty}\rho^n = P_0 \cdot \frac{1}{1-\rho} = 1$$

得

$$\boxed{\begin{array}{ll} P_0=1-\rho & \\ P_n=(1-\rho)\rho^n, \quad n\geqslant 1 & \rho<1 \end{array}} \tag{13-19}$$

这就是系统状态为 n 的概率。

上式的 ρ 有其实际意义。根据表达式的不同,可以有不同的解释。当表示为 $\rho=\lambda/\mu$ 时,它是平均到达率与平均服务率之比;即在相同时区内顾客到达的平均数与被服务的平均数之比。若表示为 $\rho=(1/\mu)/(1/\lambda)$,它是为一个顾客的服务时间与到达间隔时间之比;称 ρ 为服务强度(traffic intensity),或称 ρ 为话务强度。这是因为早期排队论是爱尔朗等人在研究电话理论时用的术语,一直沿用至今。由式(13-19),$\rho=1-P_0$,它刻画了服务机构的繁忙程度;所以又称服务机构的利用率。读者可考虑由于 ρ 的大小不同值,将会产生顾客与服务员之间、服务员与管理员之间怎样不同的反应或矛盾。

以式(13-19)为基础,可以算出系统的运行指标。

(1) 在系统中的平均顾客数(队长期望值)

$$\begin{aligned} L_s &= \sum_{n=0}^{\infty} nP_n = \sum_{n=1}^{\infty} n(1-\rho)\rho^n \\ &= (\rho+2\rho^2+3\rho^3+\cdots)-(\rho^2+2\rho^3+3\rho^4+\cdots) \\ &= \rho+\rho^2+\rho^3+\cdots = \frac{\rho}{1-\rho}, \qquad 0<\rho<1 \end{aligned}$$

或

$$L_s = \frac{\lambda}{\mu-\lambda}$$

(2) 在队列中等待的平均顾客数(队列长期望值)

$$\begin{aligned} L_q &= \sum_{n=1}^{\infty}(n-1)P_n = \sum_{n=1}^{\infty} nP_n - \sum_{n=1}^{\infty} P_n \\ &= L_s-\rho = \frac{\rho^2}{1-\rho} = \frac{\rho\lambda}{\mu-\lambda} \end{aligned}$$

关于顾客在系统中逗留的时间 W(随机变量),在 $M/M/1$ 情形下,它服从参数为 $\mu-\lambda$

的负指数分布①。

事实上，设一位顾客到达时，系统已有 n 个顾客，按先到先服务的规则，该顾客在系统中的逗留时间

$$W_n = T_1' + T_2 + \cdots + T_{n+1}$$

式中：T_1' 为正被服务的第 1 位顾客还需要的服务时间，T_i 为第 $i(i=2,\cdots,n)$ 位顾客的被服务时间，T_{n+1} 为该顾客的被服务时间。

由于 $T_i(i=2,\cdots,n+1)$ 独立且同服从参数为 μ 的负指数分布，而 T_1' 由无记忆性也服从参数为 μ 的负指数分布，从而 W_n 服从 $n+1$ 阶爱尔朗分布，即

$$f_{W_n}(t) = \frac{\mu(\mu t)^n}{n!}e^{-\mu t}$$

于是 W 的分布函数

$$F(t) = P(W \leqslant t) = \sum_{n=0}^{\infty} P_n P(W \leqslant t \mid n) = \sum_{n=0}^{\infty} \rho^n (1-\rho)\int_0^t \frac{\mu(\mu t)^n}{n!}e^{-\mu t}\,dt$$
$$= 1 - e^{-(\mu-\lambda)t} \tag{12-20}$$

即 W 服从参数为 $\mu-\lambda$ 的负指数分布，其概率密度

$$f(t) = (\mu-\lambda)e^{-(\mu-\lambda)t}$$

于是得

(3) 在系统中顾客逗留时间的期望值

$$W_s = E[W] = \frac{1}{\mu-\lambda}$$

(4) 在队列中顾客等待时间的期望值

$$W_q = W_s - \frac{1}{\mu} = \frac{\rho}{\mu-\lambda}$$

---

① 设一顾客到达时，系统已有 n 个顾客，按先到先服务的规则，这个顾客的逗留时间 W_n 就是原有各顾客的服务时间 T_i 和这个顾客服务时间 T_{n+1} 之和

$$W_n = T_1' + T_2 + \cdots + T_n + T_{n+1}$$

其中第一个顾客正被服务，T_1' 是到服务完了的部分服务时间。

令 $f(w|n+1)$ 表示 W_n 的概率密度，这是在系统已有 n 个顾客条件下的条件概率密度，所以 W 的概率密度

$$f(w) = \sum_{n=0}^{\infty} P_n f(w \mid n+1)$$

现若 $T_i(i=2,\cdots,n+1)$ 都服从参数为 μ 的负指数分布，根据负指数分布的无记忆性。T_1' 也服从同分布的负指数分布。由式(13-13)得 W_n 服从爱尔朗分布

$$f(w|n+1) = \frac{\mu(\mu w)^n e^{-\mu w}}{n!}$$

所以

$$f(w) = \sum_{n=0}^{\infty}(1-\rho)\rho^n \cdot \frac{\mu(\mu w)^n}{n!}e^{-\mu w} = (1-\rho)\mu e^{-\mu w}\sum_{n=0}^{\infty}\frac{(\rho\mu w)^n}{n!}$$
$$= (\mu-\lambda)e^{-(\mu-\lambda)w}$$

证毕。

现将以上各式归纳如下

$$\boxed{\begin{array}{ll}(1)\ L_s=\dfrac{\lambda}{\mu-\lambda} & (2)\ L_q=\dfrac{\rho\lambda}{\mu-\lambda}\\[2ex] (3)\ W_s=\dfrac{1}{\mu-\lambda} & (4)\ W_q=\dfrac{\rho}{\mu-\lambda}\end{array}} \tag{13-21}$$

它们相互的关系如下：

$$\boxed{\begin{array}{ll}(1)\ L_s=\lambda W_s & (2)\ L_q=\lambda W_q\\[2ex] (3)\ W_s=W_q+\dfrac{1}{\mu} & (4)\ L_s=L_q+\dfrac{\lambda}{\mu}\end{array}} \tag{13-22}$$

上式称为 Little 公式。

例 3 某医院手术室根据病人来诊和完成手术时间的记录，任意抽查 100 个工作小时，每小时来就诊的病人数 n 的出现次数如表 13-9 所示。又任意抽查了 100 个完成手术的病历，所用时间 v(小时)出现的次数如表 13-10 所示。

表 13-9

到达的病人数 n	出现次数 f_n
0	10
1	28
2	29
3	16
4	10
5	6
6 以上	1
合 计	100

表 13-10

为病人完成手术时间 v/小时	出现次数 f_v
0.0～0.2	38
0.2～0.4	25
0.4～0.6	17
0.6～0.8	9
0.8～1.0	6
1.0～1.2	5
1.2 以上	0
合 计	100

(1) 算出每小时病人平均到达率$=\dfrac{\sum nf_n}{100}=2.1$(人/小时)

$$\text{每次手术平均时间}=\frac{\sum vf_v}{100}=0.4(\text{小时/人})$$

每小时完成手术人数(平均服务率)$=\dfrac{1}{0.4}=2.5$(人/小时)

(2) 取 $\lambda=2.1$，$\mu=2.5$，可以通过统计检验的方法(例如 χ^2 检验法)，认为病人到达数服从参数为 2.1 的泊松分布，手术时间服从参数为 2.5 的负指数分布。

(3) $\rho=\dfrac{\lambda}{\mu}=\dfrac{2.1}{2.5}=0.84$

它说明服务机构(手术室)有 84%的时间是繁忙(被利用)，有 16%的时间是空闲的。

(4) 依次代入式(13-21),算出各指标:

在病房中病人数(期望值) $L_s=\frac{2.1}{2.5-2.1}=5.25$(人)

排队等待病人数(期望值) $L_q=0.84\times5.25=4.41$(人)

病人在病房中逗留时间(期望值) $W_s=\frac{1}{2.5-2.1}=2.5$(小时)

病人排队等待时间(期望值) $W_q=\frac{0.84}{2.5-2.1}=2.1$(小时)

不同的服务规则(先到先服务,后到先服务,随机服务)它们的不同点主要反映在等待时间的分布函数的不同,而一些期望值是相同的。我们上面讨论的各种指标,因为都是期望值,所以这些指标的计算公式对三种服务规则都适用(但对有优先权的规则不适用)。

13.3.2 系统的容量有限制的情况($M/M/1/N/\infty$)

如果系统的最大容量为 N,对于单服务台的情形,排队等待的顾客最多为 $N-1$,在某时刻一顾客到达时,如系统中已有 N 个顾客,那么这个顾客就被拒绝进入系统(见图 13-8)。

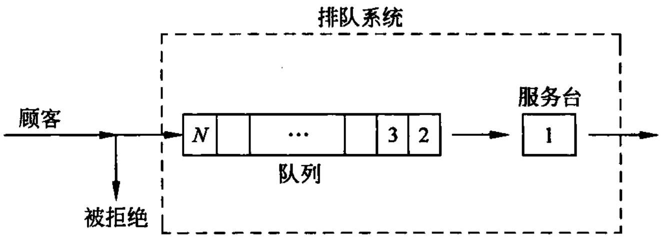


图 13-8

当 $N=1$ 时为即时制的情形;当 $N\to\infty$,为容量无限制的情形。

若只考虑稳态的情形,可作各状态间概率强度的转换关系图见图 13-9。

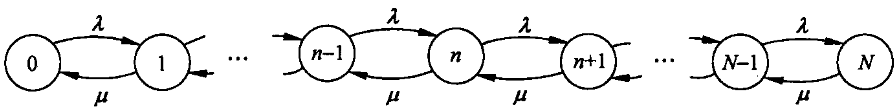


图 13-9

根据图 13-9,列出状态概率的稳态方程

$$\begin{cases}\mu P_1=\lambda P_0\\ \mu P_{n+1}+\lambda P_{n-1}=(\lambda+\mu)P_n, & n\leqslant N-1\\ \mu P_N=\lambda P_{N-1}\end{cases}\tag{13-23}$$

解这差分方程与解式(13-17)和式(13-18)是很类似的,所不同的是

$$P_0+P_1+\cdots+P_N=1$$

仍令 $\rho=\lambda/\mu$,因而得

$$\begin{cases} P_0=\dfrac{1-\rho}{1-\rho^{N+1}} & \rho\neq 1 \\ P_n=\dfrac{1-\rho}{1-\rho^{N+1}}\rho^n & n\leqslant N \end{cases} \tag{13-24}$$

这里略去 $\rho=1$ 情形的讨论(参阅习题第 13.9 题)。

在对容量没有限制的情形,我们曾设 $\rho<1$,这不仅是实际问题的需要,也是无穷级数收敛所必需的。在容量为有限数 N 的情形下,这个条件就没有必要了。(为什么?)不过当 $\rho>1$ 时,表示损失率的 P_N(或表示被拒绝排队的顾客平均数 λP_N)将是很大的。

根据式(13-24)我们可以导出系统的各种指标(计算过程略):

(1) 队长(期望值)

$$L_s=\sum_{n=0}^{N}nP_n=\frac{\rho}{1-\rho}-\frac{(N+1)\rho^{N+1}}{1-\rho^{N+1}}\quad \rho\neq 1$$

(2) 队列长(期望值)

$$L_q=\sum_{n=1}^{N}(n-1)P_n=L_s-(1-P_0)$$

当研究顾客在系统平均逗留时间 W_s 和在队列中平均等待时间 W_q 时,虽然式(13-22)仍可利用,但要注意平均到达率 λ 是在系统中有空时的平均到达率,当系统已满($n=N$)时,则到达率为 0,因此需要求出**有效到达率** $\lambda_e=\lambda(1-P_N)$。可以验证

$$1-P_0=\lambda_e/\mu$$

(3) 顾客逗留时间(期望值)

$$W_s=\frac{L_s}{\mu(1-P_0)}=\frac{L_q}{\lambda(1-P_N)}+\frac{1}{\mu}$$

(4) 顾客等待时间(期望值)

$$W_q=W_s-1/\mu$$

现在把 $M/M/1/N/\infty$ 型的指标归纳如下(当 $\rho\neq 1$ 时)

$$\begin{cases} L_s=\dfrac{\rho}{1-\rho}-\dfrac{(N+1)\rho^{N+1}}{1-\rho^{N+1}} & ① \\ L_q=L_s-(1-P_0) & ② \\ W_s=\dfrac{L_s}{\mu(1-P_0)} & ③ \\ W_q=W_s-1/\mu & ④ \end{cases} \tag{13-25}$$

例 4 单人理发馆有 6 个椅子接待人们排队等待理发。当 6 个椅子都坐满时,后来到的顾客不进店就离开。顾客平均到达率为 3 人/小时,理发需时平均 15 分钟。则

$N=7$ 为系统中最大的顾客数,$\lambda=3$ 人/小时,$\mu=4$ 人/小时。

(1) 求某顾客一到达就能理发的概率

这种情形相当于理发馆内没有顾客，所求概率

$$P_0=\frac{1-3/4}{1-(3/4)^8}=0.277\,8$$

(2) 求需要等待的顾客数的期望值

$$L_s=\frac{3/4}{1-3/4}-\frac{8(3/4)^8}{1-(3/4)^8}=2.11$$

$$L_q=L_s-(1-P_0)=2.11-(1-0.277\,8)=1.39$$

(3) 求有效到达率

$$\lambda_e=\mu(1-P_0)=4(1-0.277\,8)=2.89(\text{人/小时})$$

(4) 求一顾客在理发馆内逗留的期望时间

$$W_s=L_s/\lambda_e=2.11/2.89=0.73(\text{小时})=43.8(\text{分钟})$$

(5) 在可能到来的顾客中不等待就离开的概率($P_{n\geqslant 7}$)

这就是求系统中有 7 个顾客的概率

$$P_7=\left(\frac{\lambda}{\mu}\right)^7\left(\frac{1-\lambda/\mu}{1-(\lambda/\mu)^8}\right)=\left(\frac{3}{4}\right)^7\left[\frac{1-\frac{3}{4}}{1-\left(\frac{3}{4}\right)^8}\right]\approx 3.7\%$$

这也是理发馆的损失率。现以本例比较队长为有限和无限，两种结果如表 13-11：

表 13-11

λ=3 人/小时 μ=4 人/小时	L_s	L_q	W_s	W_q	P_0	P_7
有限队长 N=7	2.11	1.39	0.73	0.48	0.278	0.037
无限队长	3.0	2.25	1.0	0.75	0.25	0

13.3.3 顾客源为有限的情形($M/M/1/\infty/m$)

现以最常见的机器因故障停机待修的问题来说明。设共有 m 台机器(顾客总体)，机器因故障停机表示“到达”，待修的机器形成队列，修理工人是服务员，本节只讨论单服务员的情形。类似的例子还有 m 个打字员共用一台打字机，m 个会计分析员同用一个计算机终端等等。顾客总体虽只有 m 个，但每个顾客到来并经过服务后，仍回到原来总体，所以仍然可以到来。在机器故障问题中，同一台机器出了故障(到来)并经修好(服务完了)仍可再出故障(见图 13-10)。模型的符号中第 4 项，写了∞，这表示对系统的容量没有限制，但实际上它永远不会超过 m，所以和写成($M/M/1/m/m$)的意义相同。

关于平均到达率，在无限源的情形是按全体顾客来考虑的；在有限源的情形必须按每个顾客来考虑。为简单起见，设各个顾客的到达率都是相同的 λ(在这里 λ 的含义是每

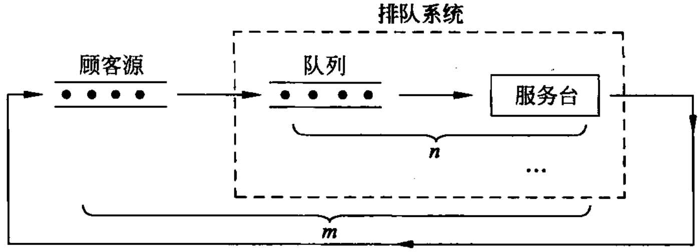


图 13-10

台机器单位运转时间内发生故障的概率或平均次数),这时在系统外的顾客平均数为 $m-L_s$,对系统的有效到达率 λ_e 应是

$$\lambda_e=\lambda(m-L_s) \tag{13-26}$$

对于$(M/M/1/\infty/m)$模型的分析可用前述的方法。在稳态的情况下,考虑状态间的转移率。当由状态 0 转移到状态 1,每台设备由正常状态转移为故障状态,其转移率为 λP_0,现有 m 台设备由无故障状态转移为有一台设备(不论哪一台)发生故障,其转移率为 $m\lambda P_0$。至于由状态 1 转移到状态 0,其状态转移率为 μP_1。所以在状态 0 时有平衡方程 $m\lambda P_0=\mu P_1$。其关系可用图 13-11 表示,由图 13-11 可得到各状态间的转移差分方程。

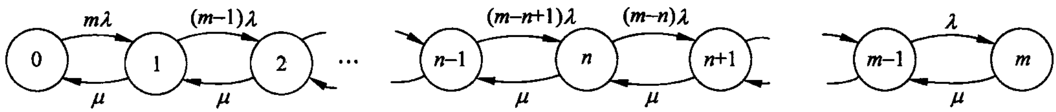


图 13-11

$$\begin{cases}\mu P_1=m\lambda P_0\\ \mu P_{n+1}+(m-n+1)\lambda P_{n-1}=[(m-n)\lambda+\mu]P_n, \quad 1\leqslant n\leqslant m-1\\ \mu P_m=\lambda P_{m-1}\end{cases}$$

解这差分方程,用递推方法,并注意到

$$\sum_{i=0}^{m}P_i=1(\text{因而不要求 }\lambda/\mu<1)$$

得

$$\begin{cases}P_0=\dfrac{1}{\sum\limits_{i=0}^{m}\dfrac{m!}{(m-i)!}\left(\dfrac{\lambda}{\mu}\right)^i}\\ P_n=\dfrac{m!}{(m-n)!}\left(\dfrac{\lambda}{\mu}\right)^n P_0 \qquad (1\leqslant n\leqslant m)\end{cases} \tag{13-27}$$

求得系统的各项指标为

$$\begin{cases} L_s = m - \dfrac{\mu}{\lambda}(1-P_0) & ① \\ L_q = m - \dfrac{(\lambda+\mu)(1-P_0)}{\lambda} = L_s - (1-P_0) & ② \\ W_s = \dfrac{m}{\mu(1-P_0)} - \dfrac{1}{\lambda} & ③ \\ W_q = W_s - 1/\mu & ④ \end{cases} \qquad (13\text{-}28)$$

在机器故障问题中 L_s 就是平均故障台数，而 $m-L_s$ 表示正常运转的平均台数。

$$m-L_s=\frac{\mu}{\lambda}(1-P_0)$$

例 5 某车间有 5 台机器，每台机器的连续运转时间服从负指数分布，平均连续运转时间 15 分钟，有一个修理工，每次修理时间服从负指数分布，平均每次 12 分钟。求：

(1) 修理工空闲的概率；

(2) 五台机器都出故障的概率；

(3) 出故障的平均台数；

(4) 等待修理的平均台数；

(5) 平均停工时间；

(6) 平均等待修理时间；

(7) 评价这些结果。

解 $m=5, \lambda=1/15, \mu=1/12, \lambda/\mu=0.8$

(1) $P_0=\left[\dfrac{5!}{5!}(0.8)^0+\dfrac{5!}{4!}(0.8)^1+\dfrac{5!}{3!}(0.8)^2+\dfrac{5!}{2!}(0.8)^3+\dfrac{5!}{1!}(0.8)^4+\dfrac{5!}{0!}(0.8)^5\right]^{-1}$

$=1/136.8=0.007\,3$

(2) $P_5=\dfrac{5!}{0!}(0.8)^5P_0=0.287$

(3) $L_s=5-\dfrac{1}{0.8}(1-0.007\,3)=3.76$(台)

(4) $L_q=3.76-0.993=2.77$(台)

(5) $W_6=\dfrac{5}{\dfrac{1}{12}(1-0.007)}-15=46$(分钟)

(6) $W_q=46-12=34$(分钟)

(7) 机器停工时间过长，修理工几乎没有空闲时间，应当提高服务率减少修理时间或增加修理工人。

13.4 多服务台负指数分布排队系统的分析

现在讨论单队、并列的多服务台(服务台数 c)的情形,我们分以下三种情形讨论。

(1) 标准的 $M/M/c$ 模型($M/M/c/\infty/\infty$);

(2) 系统容量有限制($M/M/c/N/\infty$);

(3) 有限顾客源($M/M/c/\infty/m$)。

13.4.1 标准的 $M/M/c$ 模型($M/M/c/\infty/\infty$)

关于标准的 $M/M/c$ 模型各种特征的规定与标准的 $M/M/1$ 模型的规定相同。另外规定各服务台工作是相互独立(不搞协作)且平均服务率相同 $\mu_1=\mu_2=\cdots=\mu_c=\mu$。于是整个服务机构的平均服务率为 $c\mu$(当 $n\geqslant c$);为 $n\mu$(当 $n<c$)。令 $\rho=\dfrac{\lambda}{c\mu}$,只有当$\dfrac{\lambda}{c\mu}<1$ 时才不会排成无限的队列,称它为这个系统的**服务强度或称服务机构的平均利用率**(见图 13-12)。

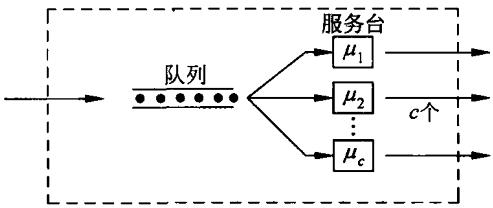


图 13-12

在分析这排队系统时,仍从状态间的转移关系开始,可见图 13-13。如状态 1 转移到状态 0,即系统中有一名顾客被服务完了(离去)的转移率为 μP_1。状态 2 转移到状态 1 时,这就是在两个服务台上被服务的顾客中有一个被服务完成而离去。因为不限哪一个,那么这时状态的转移率便是 $2\mu P_2$。同理,再考虑状态 n 转移到 $n-1$ 的情况。当 $n\leqslant c$ 时,状态转移率为 $n\mu P_n$;当 $n>c$ 时,因为只有 c 个服务台,最多有 c 个顾客在被服务,$n-c$ 个顾客在等候,因此这时状态转移率应为 $c\mu P_n$。

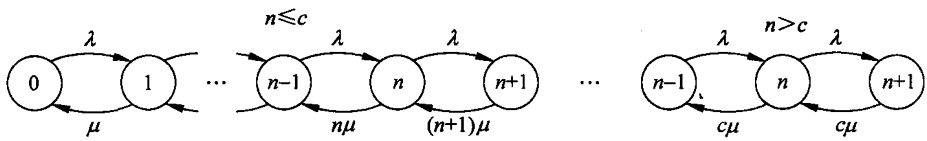


图 13-13

由图 13-13 可得

$$\begin{cases}\mu P_1=\lambda P_0 \\ (n+1)\mu P_{n+1}+\lambda P_{n-1}=(\lambda+n\mu)P_n & (1\leqslant n\leqslant c) \\ c\mu P_{n+1}+\lambda P_{n-1}=(\lambda+c\mu)P_n & (n>c)\end{cases}$$

这里 $\sum\limits_{i=0}^{\infty}P_i=1$,且 $\rho\leqslant 1$。

用递推法解上述差分方程,可求得状态概率

$$\begin{cases} P_0 = \left[\sum_{k=0}^{c-1} \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k + \frac{1}{c!} \cdot \frac{1}{1-\rho} \cdot \left(\frac{\lambda}{\mu}\right)^c\right]^{-1} \\ P_n = \begin{cases} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n P_0 & (n \leqslant c) \\ \frac{1}{c!\ c^{n-c}}\left(\frac{\lambda}{\mu}\right)^n P_0 & (n > c) \end{cases} \end{cases} \tag{13-29}$$

系统的运行指标求得如下

平均队长

$$\begin{cases} L_s = L_q + \frac{\lambda}{\mu} \\ L_q = \sum_{n=c+1}^{\infty} (n-c) P_n = \frac{(c\rho)^c \rho}{c!(1-\rho)^2} P_0 \end{cases} \tag{13-30}$$

$$\left(\text{因为} \sum_{n=c+1}^{\infty} (n-c) P_n = \sum_{n'=1}^{\infty} n' P_{n'+c} = \sum_{n'=1}^{\infty} \frac{n'}{c!c^{n'}} (c\rho)^{n'+c} P_0 = \text{右边}\right)$$

平均等待时间和逗留时间仍由 Little 公式求得

$$W_q = \frac{L_q}{\lambda}, \quad W_s = \frac{L_s}{\lambda}$$

例 6 某售票处有三个窗口,顾客的到达服从泊松过程,平均到达率每分钟 $\lambda = 0.9$ 人,服务(售票)时间服从负指数分布,平均服务率每分钟 $\mu = 0.4$ 人。现设顾客到达后排成一队,依次向空闲的窗口购票如图 13-14(a)所示,这就是一个 $M/M/c$ 型的系统,其中 $c=3$,$\frac{\lambda}{\mu}=2.25$,$\rho=\frac{\lambda}{c\mu}=\frac{2.25}{3}(<1)$符合要求的条件,代入公式得:

(1) 整个售票处空闲概率

$$P_0 = \frac{1}{\frac{(2.25)^0}{0!} + \frac{(2.25)^1}{1!} + \frac{(2.25)^2}{2!} + \frac{(2.25)^3}{3!} \cdot \frac{1}{1-2.25/3}} = 0.0748$$

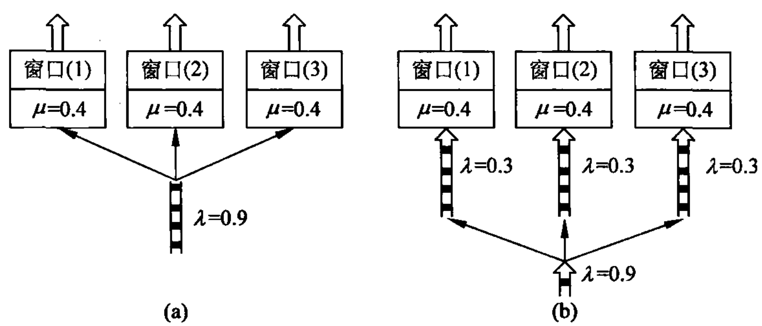


图 13-14

(2) 平均队长

$$L_q=\frac{(2.25)^3 \cdot 3/4}{3!\ (1/4)^2}\times 0.0748=1.70$$

$$L_s=L_q+\lambda/\mu=3.95$$

(3) 平均等待时间和逗留时间

$$W_q=1.70/0.9=1.89(\text{分钟})$$

$$W_s=1.89+1/0.4=4.39(\text{分钟})$$

顾客到达后必须等待(即系统中顾客数已有 3 人即各服务台都没有空闲)的概率

$$P(n\geqslant 3)=\frac{(2.25)^3}{3!\ 1/4}\times 0.0748=0.57$$

13.4.2 $M/M/c$ 型系统和 c 个 $M/M/1$ 型系统的比较

现就上面的例子说明，如果原题除排队方式外其他条件不变，但顾客到达后在每个窗口前各排一队，且进入队列后坚持不换，这就形成 3 个队列，见图 13-14(b)，而每个队列平均到达率为

$$\lambda_1=\lambda_2=\lambda_3=0.9/3=0.3(\text{每分钟})$$

这样，原来的系统就变成 3 个 $M/M/1$ 型的子系统。

现按 $M/M/1$ 型解决这个问题，并与上面比较如表 13-12 所示。

表 13-12

指 标	模 型	
	(1) $M/M/3$ 型	(2) $M/M/1$ 型
服务台空闲的概率 P_0	0.0748	0.25(每个子系统)
顾客必须等待的概率	$P(n\geqslant 3)=0.57$	0.75
平均队列长 L_q	1.70	2.25(每个子系统)
平均队长 L_s	3.95	9.00(整个系统)
平均逗留时间 W_s/分钟	4.39	10
平均等待时间 W_q/分钟	1.89	7.5

从表 13-12 中各指标的对比可以看出(1)(单队)比(2)(三队)有显著优越性，在安排排队方式时应该注意。

由于计算 P_0 和各项指标公式式(13-29)、式(13-30)很复杂，现已有专门的数值表可供使用。在式(13-29)、式(13-30)各式中 P_0 和 L_q 都是由 c 和 ρ 完全确定的，于是 $W_q\mu$ 也由 c 和 ρ 完全确定。表 13-13 给出了多服务台 $W_q\mu$ 的数值表。

表 13-13

$\lambda/c\mu$	服务台数				
	$c=1$	$c=2$	$c=3$	$c=4$	$c=5$
0.1	0.111 1	0.010 1	0.001 4	0.000 2	0.000 0*
0.2	0.250 0	0.041 7	0.010 3	0.003 0	0.001 0
0.3	0.428 6	0.098 9	0.033 3	0.013 2	0.005 8
0.4	0.666 7	0.190 5	0.078 4	0.037 8	0.019 9
0.5	1.000 0	0.333 3	0.157 9	0.087 0	0.052 1
0.6	1.500 0	0.562 5	0.295 6	0.179 4	0.118 1
0.7	2.333 3	0.960 8	0.547 0	0.357 2	0.251 9
0.8	4.000 0	1.777 8	1.078 7	0.745 5	0.554 1
0.9	9.000 0	4.263 2	2.723 5	1.969 4	1.525 0
0.95	19.000 0	9.256 4	6.046 7	4.457 1	3.511 2

* 小于 0.000 05

在例 6 中，已知 $c=3,\rho=\frac{\lambda}{c\mu}=0.75$。查表 13-13，无此数。故用线性插值法求得

$$W_q\mu=0.8129$$

因 $\mu=0.4$，所以 $W_q=2.03$ 分钟

$$W_s=2.03+\frac{1}{0.4}=4.53(\text{分钟})$$

$$L_q=2.03/0.9=2.2(\text{人})$$

$$L_s=2.2+2.25=4.45(\text{人})$$

这结果和前面计算的有差异，这是由插值引起的。

13.4.3 系统的容量有限制的情形($M/M/c/N/\infty$)

设系统的容量最大限制为 $N(\geqslant c)$，当系统中顾客数 n 已达到 N(即队列中顾客数已达 $N-c$)时，再来的顾客即被拒绝，其他条件与标准的 $M/M/c$ 型相同。

这时系统的状态概率和运行指标如下

$$P_0=\frac{1}{\sum_{k=0}^{c}\frac{(c\rho)^k}{k!}+\frac{c^c}{c!}\cdot\frac{\rho(\rho^c-\rho^N)}{1-\rho}}\qquad(\rho\neq 1)$$

$$P_n=\begin{cases}\frac{(c\rho)^n}{n!}P_0 & (0\leqslant n\leqslant c)\\ \frac{c^c}{c!}\rho^n P_0 & (c\leqslant n\leqslant N)\end{cases}\tag{13-31}$$

其中 $\rho=\frac{\lambda}{c\mu}$，但现在已不必对 ρ 加以限制(关于 $\rho=1$ 的情形可参照习题 13.9 题进行讨论)。

$$\begin{cases} L_q=\dfrac{P_0\rho(c\rho)^c}{c!\ (1-\rho)^2}[1-\rho^{N-c}-(N-c)\rho^{N-c}(1-\rho)] \\ L_s=L_q+c\rho(1-P_N) \\ W_q=\dfrac{L_q}{\lambda(1-P_N)} \\ W_s=W_q+\dfrac{1}{\mu} \end{cases} \tag{13-32}$$

由于公式的复杂,现在已有一些专门图表可供使用。

特别当 $N=c$(即时制)的情形,例如在街头的停车场就不允许排队等待空位,这时

$$\begin{cases} P_0=\dfrac{1}{\sum\limits_{k=0}^{c}\dfrac{(c\rho)^k}{k!}} \\ P_n=\dfrac{(c\rho)^n}{n!}P_0,1\leqslant n\leqslant c \end{cases} \tag{13-33}$$

其中,当 $n=c$ 即关于 P_c 的公式,被称为爱尔朗损失公式,是 A. K. Erlang 早在 1917 年发现的,并广泛应用于电话系统的设计中。

这时的运行指标如下

$$\begin{cases} L_q=0,W_q=0,W_s=\dfrac{1}{\mu} \\ L_s=\sum\limits_{n=1}^{c}nP_n=\dfrac{c\rho\sum\limits_{n=0}^{c-1}\dfrac{(c\rho)^{n-1}}{n!}}{\sum\limits_{n=0}^{c}\dfrac{(c\rho)^n}{n!}}=c\rho(1-P_c) \end{cases} \tag{13-34}$$

它又是使用的服务台数(期望值)。

例 7 在某风景区准备建造旅馆,顾客到达为泊松流,每天平均到(λ)6 人,顾客平均逗留时间($1/\mu$)为 2 天,试就该旅馆在具有(c)1,2,3,…,8 个房间的条件下,分别计算每天客房平均占用数 L_s 及满员概率 P_c。

这是即时式,因为在客房满员条件下,旅客显然不能排队等待。计算过程通过表 13-14 进行($\lambda=6,1/\mu=2,c\rho=\lambda/\mu=12$)。

第(4)栏:(2)/(3)

第(5)栏:第(4)栏各数累加

第(6)栏:(4)/(5)得满足概率 P_c,注意第(5)、(6)两栏的 c 就是同行的 n,P_c 的具体意义是:

当 $c=1$ 旅馆只有一个房间,满员(旅客被拒绝)概率 0.92

当 $c=5$ 旅馆备有 5 个房间,满员(旅客被拒绝)概率 0.63

当 $c=8$ 旅馆备有 8 个房间,满员(旅客被拒绝)概率 0.42

表 13-14

(1) n	(2) $(c\rho)^n=12^n$	(3) $n!$	(4) $(c\rho)^n/n!$	(5) $\sum_{n=0}^{c}\frac{(c\rho)^n}{n!}$	(6) P_c(答)	(7) $\sum_{n=0}^{c-1}/\sum_{n=0}^{c}$	(8) L_s(答)
0	1	1	1	1	1	—	—
1	1.2×10	1	12	13	0.92	0.08	0.92
2	1.44×10^2	2	72	85	0.85	0.15	1.83
3	1.73×10^3	6	288	373	0.77	0.23	2.74
4	2.07×10^4	24	864	1.24×10^3	0.70	0.30	3.62
5	2.49×10^5	120	2.07×10^3	3.31×10^3	0.63	0.37	4.48
6	2.99×10^6	720	4.15×10^3	7.46×10^3	0.56	0.44	5.33
7	3.58×10^7	5.04×10^3	7.11×10^3	1.45×10^4	0.49	0.51	6.14
8	4.30×10^8	4.03×10^4	1.07×10^4	2.52×10^5	0.42	0.58	6.93

第(7)栏：为求 L_s 作准备，用第(5)栏同行去除上一行结果

第(8)栏：(7)×12 得 L_s，为每天客房平均占用数，它的具体意义是：

当 $n=1$ 旅馆只有一个房间，每天客房平均占用数 $L_s=0.93$(间)

$n=5$ 旅馆备有五个房间，$L_s=4.48$(间)

$n=8$ 旅馆备有八个房间，$L_s=6.92$(间)

就是说每天平均都有一间以上的房间是空闲的。

13.4.4 顾客源为有限的情形($M/M/c/\infty/m$)

设顾客总体(顾客源)为有限数 m，且 $m>c$，和单服务台情形一样，顾客到达率 λ 是按每个顾客来考虑的，在机器管理问题中，就是共有 m 台机器，有 c 个修理工人，顾客到达就是机器出了故障，而每个顾客的到达率 λ 是指每台机器每单位运转时间出故障的期望次数。系统中顾客数 n 就是出故障的机器台数，当 $n\leqslant c$ 时，所有的故障机器都在被修理，有$(c-n)$个修理工人在空闲；当 $c<n\leqslant m$ 时，有$(n-c)$台机器在停机等待修理，而修理工人都在繁忙状态。假定这 c 个工人修理技术相同，修理(服务)时间都服从参数为 μ 的负指数分布，并假定故障的修复时间和正在生产的机器是否发生故障是相互独立的。

(1) $$P_0=\frac{1}{m!}\cdot\frac{1}{\sum_{k=0}^{c}\frac{1}{k!(m-k)!}\left(\frac{c\rho}{m}\right)^k+\frac{c^c}{c!}\sum_{k=c+1}^{m}\frac{1}{(m-k)!}\left(\frac{\rho}{m}\right)^k}$$

其中 $$\rho=\frac{m\lambda}{c\mu}$$

$$P_n=\begin{cases}\frac{m!}{(m-n)!\ n!}\left(\frac{\lambda}{\mu}\right)^n P_0 & (1\leqslant n\leqslant c)\\ \frac{m!}{(m-n)!\ c!\ c^{n-c}}\left(\frac{\lambda}{\mu}\right)^n P_0 & (c\leqslant n\leqslant m)\end{cases}\tag{13-35}$$

(2) 平均顾客数(即平均故障台数)

$$L_s=\sum_{n=1}^{m}nP_n$$

$$L_q=\sum_{n=c+1}^{m}(n-c)P_n$$

有效的到达率 λ_e 应等于每个顾客的到达率 λ 乘以在系统外(即正常生产的)机器的期望数：

$$\lambda_e=\lambda(m-L_s)$$

在机器故障问题中，它是每单位时间 m 台机器平均出现故障的次数。

(3) 可以证明

$$\begin{cases}L_s=L_q+\dfrac{\lambda_e}{\mu}=L_q+\dfrac{\lambda}{\mu}(m-L_s)\\ W_s=L_s/\lambda_e\\ W_q=L_q/\lambda_e\end{cases} \tag{13-36}$$

由于 P_0，P_n 计算公式过于复杂，有专用数值表格可供使用。

例 8 设有两个修理工人，负责 5 台机器的正常运行，每台机器平均损坏率为每运转小时 1 次，两工人能以相同的平均修复率 4 次/小时修好机器。求：

① 等待修理的机器平均数；

② 需要修理的机器平均数；

③ 有效损坏率；

④ 等待修理时间；

⑤ 停工时间。

解 $m=5,\lambda=1$(次/小时)，$\mu=4$(台/小时)，$c=2,c\rho/m=\lambda/\mu=1/4$

$$P_0=\frac{1}{5!}\left[\frac{1}{5!}\left(\frac{1}{4}\right)^0+\frac{1}{4!}\left(\frac{1}{4}\right)^1+\frac{1}{2!3!}\left(\frac{1}{4}\right)^2+\frac{2^2}{2!}\frac{1}{2!}\left(\frac{1}{8}\right)^3+\left(\frac{1}{8}\right)^4+\left(\frac{1}{8}\right)^5\right]^{-1}$$

$$=0.3149$$

$P_1=0.394,P_2=0.197,P_3=0.074,P_4=0.018,P_5=0.002$

① $L_q=P_3+2P_4+3P_5=0.118$

② $L_s=\sum_{n=1}^{m}nP_n=L_q+c-2P_0-P_1=1.094$

③ $\lambda_e=1\times(5-1.094)=3.906$

④ $W_q=0.118/3.906=0.03$(小时)

⑤ $W_s=1.094/3.906=0.28$(小时)

13.5 一般服务时间 $M/G/1$ 模型

前面我们研究了泊松输入和负指数的服务时间的模型。下面将讨论服务时间是任意分布的情形，当然，对任何情形下面关系都是正确的。

E[系统中顾客数]=E[队列中顾客数]+E[服务机构中顾客数]

E[在系统中逗留时间]=E[排队等候时间]+E[服务时间]

其中 $E[\cdot]$表示求期望值，用符号表示：

$$\begin{cases} L_s = L_q + L_{se} \\ W_s = W_q + E[T] \end{cases} \tag{13-37}$$

其中 T 表示服务时间(随机变量)，当 T 服从负指数分布时，$E[T]=1/\mu$，是讨论过的。又式(13-22)中的关系式：

$$L_s = \lambda W_s, \quad L_q = \lambda W_q$$

也是常被利用的。所以上面的 7 个指标中只要知道 3 个就可求出其余，不过在有限源和队长有限制情况下，λ 要换成有效到达率 λ_e。

13.5.1 Pollaczek-Khintchine(P-K)公式

对于 $M/G/1$ 模型，服务时间 T 的分布是一般的(但要求期望值 $E[T]$和方差$\text{Var}[T]$都存在)，其他条件和标准的 $M/M/1$ 型相同。为了达到稳态，$\rho<1$ 这一条件还是必要的，其中 $\rho=\lambda E[T]$。

在上述条件下，则有

$$L_s = \rho + \frac{\rho^2 + \lambda^2 \text{Var}[T]}{2(1-\rho)} \tag{13-38}$$

这就是 Pollaczek-Khintchine(P-K)公式。只要知道 λ，$E[T]$和 $\text{Var}[T]$，不管 T 是什么具体分布，就可求出 L_s，然后通过式(13-37)和式(13-22)可求出 L_q，W_q 和 W_s。

由式(13-38)还可知，因为有方差项的存在，在研究各期望值(各运行指标都是期望值)时，完全不考虑概率性质会得出错误结果，仅当 $\text{Var}[T]=0$ 时，随机性的波动才不影响 L_s，所以要想改进各指标，除考虑期望值外，还可以从改变方差来考虑。

现在举例说明公式的应用。

例 9 有一售票口，已知顾客按平均为 2 分 30 秒的时间间隔的负指数分布到达。顾客在售票口前服务时间平均为 2 分钟。

① 若服务时间也服从负指数分布，求顾客为购票所需的平均逗留时间和等待时间；

② 若经过调查，顾客在售票口前至少要占用 1 分钟，且认为服务时间服从负指数分布是不恰当的，而应服从以下概率密度分布，再求顾客的逗留时间和等待时间。

$$f(y)=\begin{cases}e^{-y+1} & y\geq 1\\ 0 & y<1\end{cases}$$

解 ① $\lambda=1/2.5=0.4,\mu=1/2=0.5,\rho=\lambda/\mu=0.8$

$$W_s=\frac{1}{\mu-\lambda}=10(\text{分钟})$$

$$W_q=\frac{\rho}{\mu-\lambda}=8(\text{分钟})$$

② 令 Y 为服务时间，那么 $Y=1+X$，X 服从均值为 1 的负指数分布。于是

$$E[Y]=2,\mathrm{Var}[Y]=\mathrm{Var}[1+x]=\mathrm{Var}[X]=1$$

$$\rho=\lambda E[Y]=0.8$$

代入 P-K 公式，得

$$L_s=0.8+\frac{0.8^2+0.4^2\times 1}{2\times(1-0.8)}=2.8$$

$$L_q=L_s-\rho=2$$

$$W_s=L_s/\lambda=7(\text{分钟})$$

$$W_q=L_q/\lambda=5(\text{分钟})$$

13.5.2 定长服务时间 *M/D/*1 模型

本模型的服务时间是确定的常数，例如在一条装配线上完成一件工作的时间就应是常数。自动的汽车冲洗台，冲洗一辆汽车的时间也是常数，这时

$$T=1/\mu,\mathrm{Var}[T]=0$$

$$L_s=\rho+\frac{\rho^2}{2(-\rho)} \tag{13-39}$$

例 10 某实验室有一台自动检验机器性能的仪器，要求检验机器的顾客按泊松分布到达，每小时平均 4 个顾客，检验每台机器所需时间为 6 分钟。求：

① 在检验室内机器台数 L_s(期望值，下同)；

② 等候检验的机器台数 L_q；

③ 每台机器在室内消耗(逗留)时间 W_s；

④ 每台机器平均等待检验的时间 W_q。

解 $\lambda=4,\quad E(T)=\frac{1}{10}(\text{小时}),\quad \rho=\frac{4}{10},\quad \mathrm{Var}[T]=0$

① $L_s=0.4+\frac{(0.4)^2}{2(1-0.4)}=0.533(\text{台})$

② $L_q=0.533-0.4=0.133(\text{台})$

③ $W_s=\frac{0.533}{4}=0.133(\text{小时})=8(\text{分钟})$

④ $W_q=\frac{0.133}{4}=0.033$(小时)$=2$(分钟)

注意 可以证明,在一般服务时间分布的 L_q 和 W_q 中以定长服务时间的为最小,这符合我们通俗的理解——服务时间越有规律,等候的时间就越短。读者还可在热力学或信息论中熵的概念中找出类似的性质。

13.5.3 爱尔朗服务时间 M/E$_k$/1 模型

由图 13-15 可知道,如果顾客必须经过 k 个服务站,在每个服务站的服务时间 T_i 相互独立,并服从相同的负指数分布(参数为 $k\mu$),那么 $T=\sum_{i=1}^{k}T_i$ 服从 k 阶爱尔朗分布。

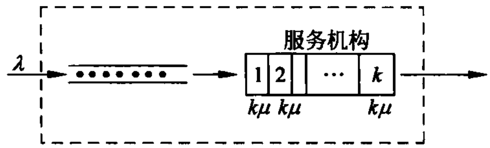


图 13-15

$$E[T_i]=\frac{1}{k\mu}\quad \mathrm{Var}[T_i]=\frac{1}{k^2\mu^2}$$

$$E[T]=\frac{1}{\mu}\quad \mathrm{Var}[T]=\frac{1}{k\mu^2}$$

对于 $M/E_k/1$ 模型(除服务时间外,其他条件与标准的 $M/M/1$ 型相同)

$$L_s=\rho+\frac{\rho^2+\frac{\lambda^2}{k\mu^2}}{2(1-\rho)}=\rho+\frac{(k+1)\rho^2}{2k(1-\rho)}$$

$$L_q=\frac{(k+1)\rho^2}{2k(1-\rho)}$$

$$W_s=L_s/\lambda,\quad W_q=L_q/\lambda \tag{13-40}$$

例 11 某单人裁缝店做西服,每套需经过 4 个不同的工序,4 个工序完成后才开始做另一套。每一工序的时间服从负指数分布,期望值为 2 小时。顾客到来服从泊松分布,平均订货率为 5.5 套/周(设一周 6 天,每天 8 小时)。问一顾客为等到做好一套西服期望时间有多长?

解 顾客到达 $\lambda=5.5$ 套/周,设:

μ——平均服务率(单位时间做完的套数);

$1/\mu$——平均每套所需的时间;

$1/4\mu$——平均每工序所需的时间。

由题设 $1/4\mu=2$(小时),$\mu=1/8$(套/小时)$=6$(套/周),$\rho=5.5/6$,设:

T_i——做完第 i 个工序所需的时间;

T——做完一套西服所需的时间。

$$E[T_i]=2,\quad \mathrm{Var}[T_i]=\left(\frac{1}{4\times 6}\right)^2$$

$$E[T]=8(\text{小时}),\quad \mathrm{Var}[T]=\frac{1}{4\times 6^2},\quad \rho=\frac{5.5}{6}$$

$$L_s=\frac{5.5}{6}+\frac{\left(\frac{5.5}{6}\right)^2+(5.5)^2\times\frac{1}{4\times 6^2}}{2\left(1-\frac{5.5}{6}\right)}=7.2188$$

顾客为等到做好一套西服的期望时间

$$W_s=L_s/\lambda=7.2188/5.5=1.3(\text{周})$$

13.6 经济分析——系统的最优化

13.6.1 排队系统的最优化问题

排队系统的最优化问题分为两类：系统设计最优化和系统控制最优化。前者称为静态问题，从排队论一诞生起就成为人们研究的内容，目的在于使设备达到最大效益，或者说，在一定的质量指标下要求机构最为经济。后者称为动态问题，是指一个给定的系统，如何运营可使某个目标函数得到最优，这是近10多年来排队论的研究重点之一。由于学习这后一问题还需更多的数学知识，所以本节只讨论静态最优的问题。

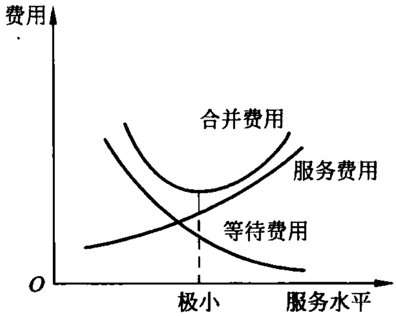


图 13-16

在一般情形下，提高服务水平(数量，质量)自然会降低顾客的等待费用(损失)，但却常常增加了服务机构的成本，我们最优化的目标之一是使二者费用之和为最小，决定达到这个目标的最优的服务水平。另一个常用的目标函数是使纯收入或使利润(服务收入与服务成本之差)为最大(见图13-16)。

各种费用在稳态情形下，都是按单位时间来考虑的。一般情形，服务费用(成本)是可以确切计算或估计的。至于顾客的等待费用就有许多不同情况，像机械故障问题中等待费用(由于机器待修而使生产遭受的损失)是可以确切估计的，但像病人就诊的等待费用(由于拖延治疗使病情恶化所受的损失)，或由于队列过长而失掉潜在顾客所造成的营业损失，就只能根据统计的经验资料来估计。

服务水平也可以由不同形式来表示，主要的是平均服务率μ(代表服务机构的服务能力和经验等)，其次是服务设备，如服务台的个数c，以及由队列所占空间大小所决定的队列最大限制数N等，服务水平也可以通过服务强度ρ来表示。

我们常用的求解方法，主要是基于求极值的思想，对于离散变量常用边际分析法，对于连续变量常用经典的微分法，对于复杂问题读者们当然可以用非线性规划或动态规划的方法。

13.6.2 $M/M/1$ 模型中最优服务率 μ

1. 标准的 $M/M/1$ 模型

取目标函数 z 为单位时间服务成本与顾客在系统逗留费用之和的期望值

$$z=c_s\mu+c_wL_s \tag{13-41}$$

其中 c_s 为当 $\mu=1$ 时服务机构单位时间的费用；c_w 为每个顾客在系统停留单位时间的费用。

将式(13-21)中 L_s 之值代入，得

$$z=c_s\mu+c_w\cdot\frac{\lambda}{\mu-\lambda}$$

为了求极小值，先求 $\frac{\mathrm{d}z}{\mathrm{d}\mu}$，然后令它为 0

$$\frac{\mathrm{d}z}{\mathrm{d}\mu}=c_s-c_w\lambda\cdot\frac{1}{(\mu-\lambda)^2}$$

$$c_s-c_w\lambda\cdot\frac{1}{(\mu-\lambda)^2}=0$$

解出最优的

$$\mu^*=\lambda+\sqrt{\frac{c_w}{c_s}\lambda} \tag{13-42}$$

根号前取＋号，是因为保证 $\rho<1$，$\mu>\lambda$ 的缘故。

2. 系统中顾客最大限制数为 N 的情形

在这情形下，系统中如已有 N 个顾客，则后来的顾客即被拒绝，于是：

P_N——被拒绝的概率(借用电话系统的术语，称为呼损率)；

$1-P_N$——能接受服务的概率；

$\lambda(1-P_N)$——单位时间实际进入服务机构顾客的平均数。在稳定状态下，它也等于单位时间内实际服务完成的平均顾客数。

设每服务 1 人能收入 G 元，于是单位时间收入的期望值是 $\lambda(1-P_N)G$ 元。

纯利润

$$\begin{aligned} z&=\lambda(1-P_N)G-c_s\mu \\ &=\lambda G\cdot\frac{1-\rho^N}{1-\rho^{N+1}}-c_s\mu \\ &=\lambda\mu G\cdot\frac{\mu^N-\lambda^N}{\mu^{N+1}-\lambda^{N+1}}-c_s\mu \end{aligned}$$

求 $\frac{\mathrm{d}z}{\mathrm{d}\mu}$，并令 $\frac{\mathrm{d}z}{\mathrm{d}\mu}=0$ 得

$$\rho^{N+1}\cdot\frac{N-(N+1)\rho+\rho^{N+1}}{(1-\rho^{N+1})^2}=\frac{c_s}{G}$$

最优的解 μ^* 应合于上式。上式中 c_s、G、λ、N 都是给定的，但要由上式中解出 μ^* 是很

困难的。通常是通过数值计算来求 μ^* 的，或将上式左方(对一定的 N)作为 ρ 的函数作出图形(见图 13-17)，对于给定的 G/c_s，根据图形可求出 μ^*/λ。

3. 顾客源为有限的情形

仍按机械故障问题来考虑。设共有机器 m 台，各台连续运转时间服从负指数分布。有 1 个修理工人，修理时间服从负指数分布。当服务率 $\mu=1$ 时的修理费用为 c_s，单位时间每台机器运转可得收入 G 元。平均运转台数为 $m-L_s$，所以单位时间纯利润为

$$z=(m-L_s)G-c_s\mu$$

$$=\frac{mG}{\rho}\cdot\frac{E_{m-1}\left(\frac{m}{\rho}\right)}{E_m\left(\frac{m}{\rho}\right)}-c_s\mu$$

式中的 $E_m(x)=\sum_{k=0}^{m}\frac{x^k}{k!}e^{-x}$ 称为泊松部分和 $\rho=\frac{m\lambda}{\mu}$，而

$$\frac{\mathrm{d}}{\mathrm{d}x}E_m(x)=E_{m-1}(x)-E_m(x)$$

为了求最优服务率 μ^*，求 $\frac{\mathrm{d}z}{\mathrm{d}\mu}$，并令 $\frac{\mathrm{d}z}{\mathrm{d}\mu}=0$，得

$$\frac{E_{m-1}\left(\frac{m}{\rho}\right)E_m\left(\frac{m}{\rho}\right)+\frac{m}{\rho}\left[E_m\left(\frac{m}{\rho}\right)E_{m-2}\left(\frac{m}{\rho}\right)-E_{m-1}^2\left(\frac{m}{\rho}\right)\right]}{E_m^2\left(\frac{m}{\rho}\right)}=\frac{c_s\lambda}{G}$$

当给定 m、G、c_s、λ，要由上式解出 μ^* 是很困难的，通常是利用泊松分布表通过数值计算来求得，或将上式左方(对一定的 m)作为 ρ 的函数作出图形(见图 13-18)，对于给定的 $\frac{c_s\lambda}{G}$ 根据图形可求出 μ^*/λ。

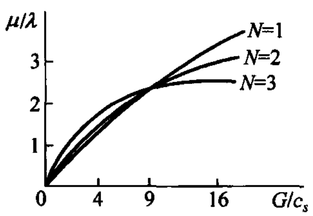


图 13-17

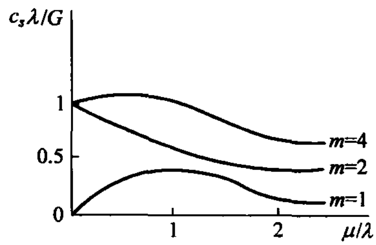


图 13-18

13.6.3 $M/M/c$ 模型中最优的服务台数 c

仅讨论标准的 $M/M/c$ 模型，且在稳态情形下，这时单位时间全部费用(服务成本与等待费用之和)的期望值

$$z=c_s' \cdot c+c_w \cdot L \tag{13-43}$$

其中 c 是服务台数；c_s' 是每服务台单位时间的成本；c_w 为每个顾客在系统停留单位时间的费用；L 是系统中顾客平均数 L_s 或队列中等待的顾客平均数 L_q（它们都随 c 值的不同而不同）。因为 c_s' 和 c_w 都是给定的，唯一可能变动的是服务台数 c，所以 z 是 c 的函数 $z(c)$，现在是求最优解 c^* 使 $z(c^*)$ 为最小。

因为 c 只取整数值，$z(c)$ 不是连续变量的函数，所以不能用经典的微分法。我们采用边际分析法（Marginal Analysis），根据 $z(c^*)$ 是最小的特点，我们有

$$\begin{cases} z(c^*)\leqslant z(c^*-1) \\ z(c^*)\leqslant z(c^*+1) \end{cases}$$

将式(13-43)中 z 代入，得

$$\begin{cases} c_s'c^*+c_wL(c^*)\leqslant c_s'(c^*-1)+c_wL(c^*-1) \\ c_s'c^*+c_wL(c^*)\leqslant c_s'(c^*+1)+c_wL(c^*+1) \end{cases}$$

上式化简后，得

$$L(c^*)-L(c^*+1)\leqslant c_s'/c_w\leqslant L(c^*-1)-L(c^*) \tag{13-44}$$

依次求 $c=1,2,3,\cdots$ 时 L 的值，并作两相邻的 L 值之差，因 c_s'/c_w 是已知数，根据这个数落在哪个不等式的区间里就可定出 c^*。

例 12 某检验中心为各工厂服务，要求做检验的工厂（顾客）的到来服从泊松流，平均到达率 λ 为每天 48 次，每次来检验由于停工等原因损失为 6 元。服务（做检验）时间服从负指数分布，平均服务率 μ 为每天 25 次，每设置 1 个检验员服务成本（工资及设备损耗）为每天 4 元。其他条件适合标准的 $M/M/c$ 模型，问应设几个检验员（及设备）才能使总费用的期望值为最小？

解 $c_s'=4$ 元/检验员　　$c_w=6$ 元/次

$$\lambda=48 \qquad \mu=25 \qquad \lambda/\mu=1.92$$

设检验员数为 c，令 c 依次为 1,2,3,4,5，根据表 13-13，求出 L_s。计算过程如表 13-15 所示。

表 13-15

c	1	2	3	4	5
$\lambda/c\mu$	1.92	0.96	0.64	0.48	0.38
查表 $W_q \cdot \mu$	—	10.255 0	0.396 1	0.077 2	0.017 0
$L_s=\frac{\lambda}{\mu}(W_q \cdot \mu+1)$	—	21.610	2.680	2.068	1.952

将 L_s 值代入式(13.44)得表 13-16。

表 13-16

检验员数 c	来检验顾客数 $L_s(c)$	$L(c)-L(c+1)\sim$ $L(c)-L(c-1)$	总费用(每天)$z(c)$
1	∞		∞
2	21.610	18.930～∞	154.94
3	2.680	0.612～18.930	27.87(*)
4	2.068	0.116～0.612	28.38
5	1.952		31.71

$\frac{c_s'}{c_w}=0.666$,落在区间(0.612～18.930)内,所以 $c^*=3$。即以设 3 个检验员使总费用为最小,直接代入式(13-43)也可验证总费用为最小。

$$z(c^*)=z(3)=27.87(\text{元})$$

13.7* 分析排队系统的随机模拟法

当排队系统的到达间隔时间和服务时间的概率分布很复杂时,或不能用公式给出时,就不能用解析法求解,这时可采用随机模拟法求解,现举例说明。

例 13 设某仓库前有一卸货场,货车一般是夜间到达,白天卸货。每天只能卸货 2 车,若一天内到达数超过 2 车,那么就推迟到次日卸货。根据表 13-17 所示的经验货车到达数的概率分布(相对频率)平均为 1.5 车/天,求每天推迟卸货的平均车数。

表 13-17

到达车数	0	1	2	3	4	5	≥6
概率	0.23	0.30	0.30	0.1	0.05	0.02	0.00

解 这是单服务台的排队系统,可验证到达车数不服从泊松分布,服务时间也不服从负指数分布(这是定长服务时间),不能用以前的方法求解。

随机模拟法首先要求事件能按历史的概率分布规律出现。对例 13 的数据进行分析,取 100 张卡片,按表 13-17 的概率,取 23 张卡片填入 0;取 30 张填 1;取 30 张填 2;取 10 张填 3;取 5 张填 4;取 2 张填 5 等。然后将这些卡片放在盒内搅均匀,再随机地一一取出,依次记录卡片上的数码,得到这一系列数据就是每天到达车数的模拟。实际应用时可用随机数表,表 13-18 就是随机数表的一部分。

也可以用 MS Excel 表中的 RAND()函数产生随机数 x,在 $0<x<1$ 之间,若需要整数两位的随机数,可以用 ROUND(RAND() * 100,0)产生随机数。再用复制、粘贴两个操作,完成需要的随机数的个数,比用随机数表方便。

表 13-18

97	95	12	11	90	49	57	13	86	81
02	92	75	91	24	58	39	22	13	02
80	67	14	99	16	89	96	63	67	60
66	24	72	57	32	15	49	63	00	04
96	76	20	28	72	12	77	23	79	46
55	64	82	61	73	94	26	18	37	31
50	02	74	70	16	85	95	32	85	67
29	53	08	33	81	34	30	21	24	25
58	16	01	91	70	07	50	13	18	24
51	16	69	67	16	53	11	06	36	10
04	55	36	97	30	99	80	10	52	40
86	54	35	61	59	89	64	97	16	02
24	23	52	11	59	10	88	68	17	39
39	36	99	50	74	27	69	48	32	68
47	44	41	86	83	50	24	51	02	08
60	71	41	25	90	93	07	24	29	59
65	88	48	06	68	92	70	97	02	66
44	74	11	60	14	57	08	54	12	90
93	10	95	80	32	50	40	44	08	12
20	46	36	19	47	78	16	90	59	64
86	54	24	88	94	14	58	49	80	79
12	88	12	25	19	70	40	06	40	31
42	00	50	24	60	90	69	60	07	86
29	98	81	68	61	24	90	92	32	68
36	63	02	37	89	40	81	77	74	82
01	77	82	78	20	72	35	38	56	89
41	69	43	37	41	21	36	39	57	80
54	40	76	04	05	01	45	84	55	11
68	03	82	32	22	80	92	47	77	62
21	31	77	75	43	13	83	43	70	16
53	64	54	21	04	23	85	44	81	36
91	66	21	47	95	69	58	91	47	59
48	72	74	40	97	92	05	01	61	18

本例在求解时先按到达车数的概率，分别给它们分配随机数，见表 13-19①。

① 严格地说，这里利用了概率论中的一个定理。若 X 的分布函数是 $F(x)$，R 是[0,1]区间上均匀分布的随机变量，那么 $X=F^{-1}(R)$，其中 $F^{-1}(\cdot)$表示反函数。利用此定理可由分布函数求到 X 的抽样值。

表 13-19

到达车数	概　　率	累积概率	对应的随机数
0	0.23	0.23	00～22
1	0.30	0.53	23～52
2	0.30	0.83	53～82
3	0.10	0.93	83～92
4	0.05	0.98	93～97
5	0.02	1.00	98～99
	1.00		

以下开始模拟(见表 13-20)。前 3 天作为模拟的预备期,记为 x。然后依次从第 1 天、第 2 天……第 50 天。如第 1 天得到随机数 66,从表 13-19 中可见,第 1 天达到车数为 2,将它记入表 13-20。第 2 天,得到随机数 96,它在表 13-19 中,对应达到 4 车……如此一直到第 50 天。表 13-20 的第(2)、(3)列数字都填入后,计算出第(4)、(5)、(6)列数字,从第一个 x 日开始。当天到车数(3)+前一天推迟车数(6)=当天需要卸货车数(4);

$$\text{卸货车数(5)}=\begin{cases}\text{需要卸货车数(4)}, & \text{当天需要卸货车数}\leqslant 2\\ 2 & \text{当天需要卸货车数}>2\end{cases}$$

分析结果时,不考虑头三天写 x 的预备阶段的数据。这是为了使模拟在一个稳态过程中任意点开始,否则若认为开始时没有积压就失去随机性了。表 13-20 中表明了模拟 50 天运行情况,这相当于一个随机样本。由此可见多数情况下很少发生推迟卸车而造成积压。只是在第 36 天比较严重,平均达车数为 1.58,比期望值略高。又知平均每天有 0.9 车推迟卸货,当然模拟时间越长结果越准确。这方法适用于对不同方案可能产生的结果进行比较,用电子计算机进行模拟更为方便。模拟方法只能得到数字结果,不能得出解析式。

表 13-20

(1)日期	(2)随机数	(3)到达数	(4)需要卸货车数	(5)卸货车数	(6)推迟卸货车数
x	97	4	4	2	2
x	02	0	2	2	0
x	80	2	2	2	0
1	66	2	2	2	0
2	96	4	4	2	2
3	55	2	4	2	2
4	50	1	3	2	1
5	29	1	2	2	0
6	58	2	2	2	0
7	51	1	1	1	0
8	04	0	0	0	0
9	86	3	3	2	1
10	24	1	2	2	0

续表

(1)日期	(2)随机数	(3)到达数	(4)需要卸货车数	(5)卸货车数	(6)推迟卸货车数
11	39	1	1	1	0
12	47	1	1	1	0
13	60	2	2	2	0
14	65	2	2	2	0
15	44	1	1	1	0
16	93	4	4	2	2
17	20	0	2	2	0
18	86	3	3	2	1
19	12	0	1	1	0
20	42	1	1	1	0
21	29	1	1	1	0
22	36	1	1	1	0
23	01	0	0	0	0
24	41	1	1	1	0
25	54	2	2	2	0
26	68	2	2	2	0
27	21	0	0	0	0
28	53	2	2	2	0
29	91	3	3	2	1
30	48	1	2	2	0
31	36	1	1	1	0
32	55	2	2	2	0
33	70	2	2	2	0
34	38	1	1	1	0
35	36	1	1	1	0
36	98	5	5	2	3
37	50	1	4	2	2
38	95	4	6	2	4
39	92	3	7	2	5
40	67	2	7	2	5
41	24	1	6	2	4
42	76	2	6	2	4
43	64	2	6	2	4
44	02	0	4	2	2
45	53	2	4	2	2
46	16	0	2	2	0
47	16	0	0	0	0
48	55	2	2	2	0
49	54	2	2	2	0
50	23	1	1	1	0
总　计		79			45
平　均		1.58			0.90

习题

13.1 某工地为了研究发放工具应设置几个窗口，对于请领和发放工具分别作了调查记录：

(1) 以 10 分钟为一段，记录了 100 段时间内每段到来请领工具的工人数，见表 13-21；

(2) 记录了 1 000 次发放工具(服务)所用时间(秒)，见表 13-22，试求：

① 平均到达率和平均服务率。

② 利用统计学的方法证明：若假设到来的数是服从参数 $\lambda=1.6$ 的泊松分布，服务时间服从参数 $\mu=0.9$ 的负指数分布，这是可以接受的。

(3) 这时只设一个服务员是不行的，为什么？试分别就服务员数 $c=2,3,4$ 各情况计算等待时间 W_q，注意用表 13-13。

(4) 设请领工具的工人等待的费用损失为每小时 6 元，发放工具的服务员空闲费用损失为每小时 3 元，每天按 8 小时计算，问设几个服务员使总费用损失为最小？

表 13-21

每 10 分钟内领工具人数	次　数
5	1
6	0
7	1
8	1
9	1
10	2
11	4
12	6
13	9
14	11
15	12
16	13
17	10
18	9
19	7
20	4
21	3
22	3
23	1
24	1
25	1
合　计	100

表 13-22

发放时间/秒	次　数
15	200
30	175
45	140
60	104
75	78
90	69
105	51
120	47
135	38
150	30
165	16
180	12
195	10
210	7
225	9
240	9
255	3
270	1
285	1
合　计	1 000

13.2 某修理店只有一个修理工人，来修理的顾客到达次数服从泊松分布，平均每小时 4 人，修理时间服从负指数分布，平均需 6 分钟。求：

(1) 修理店空闲时间概率；

(2) 店内有 3 个顾客的概率；

(3) 店内至少有 1 个顾客的概率；

(4) 在店内顾客平均数；

(5) 在店内平均逗留时间；

(6) 等待服务的顾客平均数；

(7) 平均等待修理(服务)时间；

(8) 必须在店内消耗 15 分钟以上的概率。

13.3 在某单人理发店顾客到达为泊松流，平均到达间隔为 20 分钟，理发时间服从负指数分布，平均时间为 15 分钟。求：

(1) 顾客来理发不必等待的概率；

(2) 理发店内顾客平均数；

(3) 顾客在理发店内平均逗留时间；

(4) 若顾客在店内平均逗留时间超过 1.25 小时，则店主将考虑增加设备及理发员，问平均到达率提高多少时店主才作这样考虑呢？

13.4 在例 3 中

(1) 试求系统中(包括手术室和候诊室)有 0,1,2,3,4,5 个病人的概率。

(2) 设 λ 不变而 μ 是可控制的，证明：若医院管理人员认为使病人在医院平均耗费时间超过 2 小时是不允许的，那么必须平均服务率 μ 达到 2.6(人/小时)以上。

13.5 称顾客为等待所费时间与服务时间之比为顾客损失率，用 R 表示。

(1) 试证：对于 $M/M/1$ 模型有 $R=\dfrac{\lambda}{\mu-\lambda}$；

(2) 在例 3 中仍设 λ 不变，μ 是可控制的，试定 μ 使顾客损失率小于 4。

13.6 设 n_s 表示系统中顾客数，n_q 表示队列中等候的顾客数，在单服务台系统中有

$$n_s=n_q+1(n_s,n_q>0)$$

试说明它们的期望值

$$L_s\neq L_q+1$$

$$L_s=L_q+\rho$$

根据这关系式给 ρ 以直观解释。

13.7 某工厂为职工设立了昼夜 24 小时都能看病的医疗室(按单服务台处理)。病人到达的平均间隔时间为 15 分钟，平均看病时间为 12 分钟，且服从负指数分布，因工人看病每小时给工厂造成损失为 30 元。

(1) 试求工厂每天损失的期望值。

(2) 问平均服务率提高多少，方可使上述损失减少一半？

13.8 对于 $M/M/1/\infty/\infty$ 模型，在先到先服务情况下，试证：顾客排队等待时间分布概率密度是

$$f(w_q)=\lambda(1-\rho)e^{-(\mu-\lambda)w_q}, \quad w_q>0$$

并根据上式求等待时间的期望值 W_q。

13.9 在 $M/M/1/N/\infty$ 模型中，若 $\rho=1$（即 $\lambda=\mu$），试证：式(13-24)应为

$$P_0=P_1=\cdots=\frac{1}{N+1}$$

于是

$$L_s=N/2$$

13.10 对于 $M/M/1/N/\infty$ 模型，试证：

$$\lambda(1-P_n)=\mu(1-P_0)$$

并对上式给予直观的解释。

13.11 某新工厂正在决定分配给一个特别的工作中心多少存贮空间。工作以平均每小时三个的泊松分布被送到工作中心，工作中心每次只能执行一个工作，完成该工作所需要的时间遵从每小时 0.25 个的指数分布。若工作到达时工作中心内已满，则工作需转放到一个不方便的地方。如果每个工作在工作中心存放时需要一平方米的空间，工厂希望工作中心内的空间能够保证 90%的时间里容纳下全部到达的工作，需要分配多少空间给工作中心？$\left(\text{提示：几何数列的和}\sum_{n=0}^{N}x^n=\frac{1-x^{N+1}}{1-x}\right)$

13.12 在习题 13.2 中若店内已有 3 个顾客，那么后来的顾客即不再排队，其他条件不变。试求：

(1) 店内空闲的概率；

(2) 各运行指标 L_s, L_q, W_s, W_q。

13.13 在习题 13.2 中，若顾客平均到达率增加到每小时 12 人，仍为泊松流，服务时间不变，这时增加了一个工人。

(1) 根据 λ/μ 的值说明增加工人的原因。

(2) 增加工人后求店内空闲概率；店内有 2 个或更多顾客(即工人繁忙)的概率。

(3) 求 L_s, L_q, W_s, W_q。

13.14 有 $M/M/1/5/\infty$ 模型，平均服务率 $\mu=10$，就两种到达率：$\lambda=6, \lambda=15$，已计算出相应的概率 P_n 如表 13-23 所示。试就这两种情况计算。

表 13-23

系统中顾客数 n	$(\lambda=6)P_n$	$(\lambda=15)P_n$
0	0.42	0.05
1	0.25	0.07
2	0.15	0.11
3	0.09	0.16
4	0.05	0.24
5	0.04	0.37

(1) 有效到达率和服务台的服务强度；

(2) 系统中平均顾客数；

(3) 系统的满员率；

(4) 服务台应从哪些方面改进工作？理由是什么？

13.15 对于 $M/M/1/m/m$ 模型，试证：

$$L_s = m - \frac{\mu(1-P_0)}{\lambda}$$

并给予直观解释。

13.16 对于 $M/M/c/\infty/\infty$ 模型，μ 是每个服务台的平均服务率，试证：

(1) $L_s - L_q = \lambda/\mu$；

(2) $\lambda = \mu\left[c - \sum_{n=0}^{c}(c-n)P_n\right]$。

并给予直观解释。

注意 在单服务台情况，(1)式是很容易解释的。但是 c 个服务台时，其结果仍相同，且与 c 无关，这是引人注意的。

13.17 机场有两条跑道，一条专供起飞用，一条专供降落用。已知要求起飞和降落的飞机都分别按平均每小时 25 架次的泊松流到达，每架飞机起飞或降落占用跑道的时间都服从平均 2 分钟的负指数分布。又设起飞和降落是彼此无关的。

(1) 试求一架飞机起飞或降落为等待使用跑道所需的平均时间；

(2) 若机场拟调整使用跑道办法，每条跑道都可做起飞或降落用。但为了安全，每架飞机占用跑道时间延长为平均 2.16 分钟的负指数分布，这时要求起飞和要求降落的飞机将混合成一个参数为 50 架次/小时的泊松到达流。试计算这种情形下的平均等待时间。

(3) 以上两种办法哪个更好些呢？

13.18 车间内有 m 台机器，有 c 个修理工($m>c$)。每台机器发生故障率为 λ，符合 $M/M/c/m/m$ 模型，试证：

$$\frac{W_s}{\left(\frac{1}{\lambda}\right)+W_s} = \frac{L_s}{m}$$

并说明上式左右两端的概率意义。

13.19 在例 9 中如售票处使用自动售票机，顾客在窗口前的服务时间将减少 20%。这时认为服务时间分布的概率密度是

$$f(z)=\begin{cases}1.25e^{-1.25z+1} & z\geqslant 0.8\\ 0 & z<0.8\end{cases}$$

(这里的服务时间 z 与例 9 中(2)的 y 关系很相似，$z=0.8y$)再求顾客的逗留时间和等待时间。

13.20　在第 2 题中，如服务时间服从正态分布，数学期望值仍是 6 分钟，方差 $\sigma^2=1/8$，求店内顾客数的期望值。

13.21　一个办事员核对登记的申请书时，必须依次检查 8 张表格，核对每份申请书需 1 分钟。顾客到达率为每小时 6 人，服务时间和到达间隔均为负指数分布，求：

(1) 办事员空闲的概率；

(2) L_s，L_q，W_s 和 W_q。

13.22　对于单服务台情形，试证：

(1) 定长服务队列 $L_q^{(1)}$ 是负指数服务队列 $L_q^{(2)}$ 的一半；

(2) 定长服务时间 $W_q^{(1)}$ 是负指数服务时间 $W_q^{(2)}$ 的一半。

Part 8

第 8 篇

存储论

CHAPTER 14

第14章

存 储 论

14.1 存储论的基本概念

14.1.1 存储问题的提出

人们在生产和日常生活活动中往往将所需的物资、用品和食物暂时地储存起来，以备将来使用或消费。这种储存物品的现象是为了解决供应(生产)与需求(消费)之间的不协调的一种措施，这种不协调性一般表现为供应量与需求量和供应时期与需求时期的不一致性上，出现供不应求或供过于求。人们在供应与需求这两环节之间加入储存这一环节，就能起到缓解供应与需求之间的不协调，以此为研究对象，利用运筹学的方法去解决最合理、最经济地储存问题。例如：

(1) 水电站在雨季到来之前，水库应蓄水多少？这个问题就存在一个矛盾。就发电的需要来说，当然蓄水以多为好。就安全来说，如果雨季降雨量大，则必须考虑先放掉一些水，使水库存水量减少。否则洪水到来时，水库水位猛涨，溢洪道排泄不及时，可能会使水坝坍塌，除水电站被破坏外，还会给下游造成巨大的损失。假设只考虑安全，可提前把水库存水放空。但当雨季降雨量小时，就会造成水库存水量不足，使发电量减少。因此，合理地调节水库的存水量对国民经济有重大意义。比如 1981 年第一季度，由于我国各地水电站调度合理，比计划多发电 14 亿度，相当于节约用煤 80 万吨。

(2) 工厂生产需用原料，如没有储存一定数量的原料，会发生停工待料现象。原料储存过多除积压资金外，还要支付一笔存储保管费用。如蛋品加工厂需用鲜蛋做原料，鲜蛋存量过少，会产生停工待料而使工厂开工不足而造成损失；如储存鲜蛋过多，在支付存储保管费用中还要支出一笔冷冻保鲜费用。遇到一些意外的因素使鲜蛋变质则损失更大。在这种情况下，最好存储多少鲜蛋？

(3) 在机器制造厂中，加工一个零件常常需要经过许多工序。一道工序完工后即成为下道工序的生产备件，在这种由前道工序转入下道工序的环节中就会产生存储问题。每台机器由许多部件装配而成，每个部件又由许多零件组成，每个零件又需要许多工序才能制成。因此发生在每个环节中的存储问题，总起来看就不是可以忽略不计的问题了。如果各个环节生产备件的存储量安排得合理，既可避免因停工待料造成经济上的损失，又可以少占用流动资金，加速周转，增加利润。

(4) 在商店里若存储商品数量不足，会发生缺货现象，失去销售机会而减少利润；如果存量过多，一时售不出去，会造成商品积压，占用流动资金过多而且周转不开，这样也会给商家造成经济损失。当然，顾客购买何种商品以及购买多少，都带有随机性，在这种情况下商店的管理人员就应研究商品的存储量。

诸如此类，与存储量有关的问题，需要人们作出抉择，在长期实践中人们摸索到一些规律，也积累了一些经验。专门研究这类有关存储问题的科学，构成运筹学的一个分支，叫做**存储论**(inventory)，也称库存论。

本章所介绍的存储问题，模型并不复杂，原理也容易掌握，应用这些原理可以从一个方面改善企业的经营管理，以达到节约资金，获得更多利润的目的。

14.1.2 存储论的基本概念

工厂为了生产，必须储存一些原料，把这些储存物简称**存储**。生产时从存储中取出一定数量的原料消耗掉，使存储减少。生产不断进行，存储不断减少，到一定时刻必须对存储给以补充，否则存储用完了，生产无法进行。

商店必须储存一些商品(即存储)，营业时卖掉一部分商品使存储减少，到一定的时候又必须进货，否则库存售空无法继续营业。

一般地说，存储量因需求而减少，因补充而增加。

1. 需求

对存储来说，由于需求，从存储中取出一定的数量，使存储量减少，这就是存储的输出。有的需求是间断式的，有的需求是连续均匀的。

图 14-1 和图 14-2 分别表示 t 时间内的输出量皆为 S-W，但两者的输出方式不同。图 14-1 表示输出是间断的，图 14-2 表示输出是连续的。

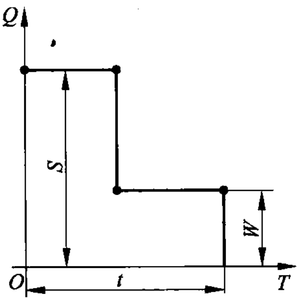


图 14-1

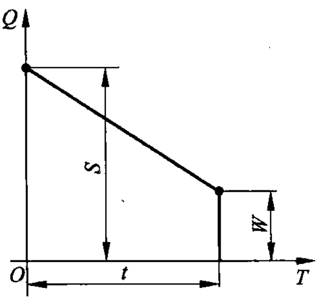


图 14-2

有的需求是确定性的，如钢厂每月按合同卖给电机厂矽钢片 10 吨。有的需求是随机性的，如书店每日卖出去的书可能是 1 000 本，也可能是 800 本。但是经过大量的统计以后，可能会发现每日售书数量的统计规律，称之为有一定的随机分布的需求。

2. 补充(订货或生产)

存储由于需求而不断减少,必须加以补充,否则最终将无法满足需求。补充就是存储的输入。补充的办法可能是向其他工厂购买,从订货到货物进入"存储"往往需要一段时间,我们把这段时间称为**备货时间**。从另一个角度看,为了在某一时刻能补充存储,必须提前订货,那么这段时间也可称之为**提前时间**(leadtime)。

备货时间可能很长,也可能很短,可能是随机性的,也可以是确定性的。

存储论要解决的问题是:多少时间补充一次,每次补充的数量应该是多少。决定多少时间补充一次以及每次补充数量的策略称为**存储策略**。

存储策略的优劣如何衡量呢?最直接的衡量标准,是计算该策略所耗用的平均费用多少。为此有必要对费用进行详细的分析。

3. 费用

主要包括下列一些费用:

(1) 存储费,包括货物占用资金应付的利息以及使用仓库、保管货物、货物损坏变质等支出的费用。

(2) 订货费,包括两项费用,一项是订购费用(固定费用)如手续费、电信往来、派人员外出采购等费用。订购费与订货次数有关而与订货数量无关。另一项是货物的成本费用,它与订货数量有关(可变费用),如货物本身的价格,运费等。如货物单价为 K 元,订购费用为 C_3 元,订货数量为 Q,则订货费用为:C_3+KQ。

(3) 生产费,补充存储时,如果不需向外厂订货,由本厂自行生产,这时仍需要支出两项费用。一项是装配费用(或称准备、结束费用,是固定费用),如更换模、夹具需要工时,或添置某些专用设备等属于这项费用,也用 C_3 表示。另一项是与生产产品的数量有关的费用如材料费、加工费等(可变费用)。

(4) 缺货费,当存储供不应求时所引起的损失。如失去销售机会的损失、停工待料的损失以及不能履行合同而缴纳罚款等。

在不允许缺货的情况下,在费用上处理的方式是缺货费为无穷大。

4. 存储策略

如前所述决定何时补充,补充多少数量的办法称之为存储策略,常见的策略有三种类型。

(1) t_0-循环策略,每隔 t_0 时间补充存储量 Q。

(2) (s,S)策略,每当存储量 $x>s$ 时不补充。当 $x\leqslant s$ 时补充存储。补充量$Q=S-x$(即将存储量补充到 S)。

(3) (t,s,S)混合策略,每经过 t 时间检查存储量 x,当 $x>s$ 时不补充。当 $x\leqslant s$ 时,补充存储量使之达到 S。

确定存储策略时,首先是把实际问题抽象为数学模型。在形成模型过程中,对一些复杂的条件尽量加以简化,只要模型能反映问题的本质就可以了。然后对模型用数学的方

法加以研究,得出数量的结论。这结论是否正确,还要拿到实践中加以检验。如结论与实际不符,则要对模型重新加以研究和修改。存储问题经长期研究已得出一些行之有效的模型。从存储模型来看大体上可分为两类:一类叫做确定性模型,即模型中的数据皆为确定的数值;另一类叫做随机性模型,即模型中含有随机变量,而不是确定的数值。

由于具体条件有差别,制定存储策略时又不能忽视这些差别,因而模型也有多种类型。本章将按确定性存储模型及随机性存储模型两大类,分别介绍一些常用的存储模型,并从中得出相应的存储策略。

一个好的存储策略,既可以使总费用小,又可避免因缺货影响生产(或对顾客失去信用),下面利用一些具体的模型阐述如何求出较好的存储策略。

14.2 确定性存储模型

14.2.1 模型一:不允许缺货,备货时间很短

在研究、建立模型时,需要做一些假设,目的是使模型简单、易于理解、便于计算。为此作如下假设:

(1) 缺货费用无穷大;

(2) 当存储降至零时,可以立即得到补充(即备货时间或拖后时间很短,可以近似地看做零);

(3) 需求是连续的、均匀的,设需求速度 R(单位时间的需求量)为常数,则 t 时间的需求量为 Rt;

(4) 每次订货量不变,订购费不变(每次备货量不变,装配费不变);

(5) 单位存储费不变。

存储量变化情况用图 14-3 表示。

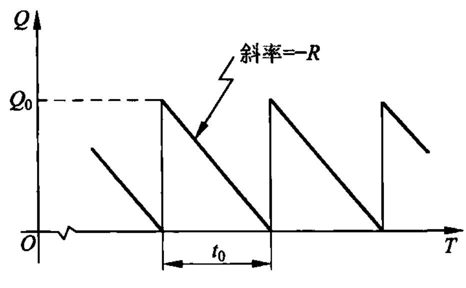


图 14-3

由于可以立即得到补充,所以不会出现缺货,在研究这种模型时不再考虑缺货费用。这些假设条件只是近似的正确,在这些假设条件下如何确定存储策略呢? 正如 14.1.2 中

所提示的，要用总平均费用来衡量存储策略的优劣。为了找出最低费用的策略，首先想到在需求确定的情况下，每次订货量多，则订货次数可以减少，从而减少了订购费。但是每次订货量多，会增加存储费用。为研究费用的变化情况需要导出费用函数。

假定每隔 t 时间补充一次存储，那么订货量必须满足 t 时间的需求 Rt，记订货量为 Q，$Q=Rt$，订购费为 C_3，货物单价为 K，则订货费为 C_3+KRt；t 时间的平均订货费为 $\frac{C_3}{t}+KR$，t 时间内的平均存储量为

$$\frac{1}{t}\int_0^t RT\,\mathrm{d}T=\frac{1}{2}Rt$$

(此结果由图 14-3 中利用几何知识易得出，平均存储量为三角形高的 1/2)单位时间内单位物品的存储费用为 C_1，t 时间内所需平均存储费用为 $\frac{1}{2}RtC_1$。

t 时间内总的平均费用为

$$C(t)=\frac{C_3}{t}+KR+\frac{1}{2}C_1Rt \tag{14-1}$$

t 取何值时 $C(t)$ 最小？只需对式(14-1)利用微积分求最小值的方法可求出。

令

$$\frac{\mathrm{d}C(t)}{\mathrm{d}t}=-\frac{C_3}{t^2}+\frac{1}{2}C_1R=0$$

得

$$t_0=\sqrt{\frac{2C_3}{C_1R}} \tag{14-2}$$

因 $\frac{\mathrm{d}^2C(t)}{\mathrm{d}t^2}>0$，即每隔 t_0 时间订货一次可使 $C(t)$ 最小。

订货批量

$$Q_0=Rt_0=\sqrt{\frac{2C_3R}{C_1}} \tag{14-3}$$

式(14-3)即存储论中著名的经济订购批量(economic ordering quantity)公式。简称为 E. O. Q 公式，也称平方根公式，或经济批量(economic lot size)公式。由于 Q_0、t_0 皆与 K 无关，所以此后在费用函数中略去 KR 这项费用。如无特殊需要不再考虑此项费用，式(14-1)改写为

$$C(t)=\frac{C_3}{t}+\frac{1}{2}C_1Rt \tag{14-4}$$

将 t_0 代入式(14-4)得出最佳费用

$$\begin{aligned}C_0&=C(t_0)=C_3\sqrt{\frac{C_1R}{2C_3}}+\frac{1}{2}C_1R\sqrt{\frac{2C_3}{C_1R}}\\&=\sqrt{2C_1C_3R}\end{aligned}$$

$$C_0=\min C(t) \tag{14-5}$$

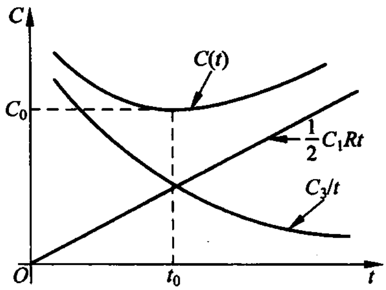


图 14-4

从费用曲线(见图 14-4)也可以求出 t_0,Q_0,C_0。

存储费用曲线

$$\frac{1}{2}C_1Rt$$

订购费用曲线

$$\frac{C_3}{t}$$

总费用曲线

$$C(t)=\frac{C_3}{t}+\frac{1}{2}C_1Rt \tag{14-1'}$$

$C(t)$曲线的最低点($\min C(t)$)的横坐标 t_0 与存储费用曲线、订购费用曲线交点横坐标相同。即

$$\frac{C_3}{t_0}=\frac{1}{2}C_1Rt_0$$

解出

$$t_0=\sqrt{\frac{2C_3}{C_1R}} \tag{14-2'}$$

$$Q_0=Rt_0=\sqrt{\frac{2C_3R}{C_1}} \tag{14-3'}$$

$$C_0=\frac{C_3}{t_0}+\frac{1}{2}C_1Rt_0=\sqrt{2C_1C_3R} \tag{14-4'}$$

式(14-2′),式(14-3′),式(14-4′)与式(14-2),式(14-3),式(14-5)一致。

式(14-2)是由于选 t 作为存储策略变量推导出来的。如果选订货批量 Q 作存储策略变量也可以推导出上述公式。

例 1 某厂按合同每年需提供 D 个产品,不许缺货。假设每一周期工厂需装配费 C_3 元,存储费每年每单位产品为 C_1 元,问全年应分几批供货才能使装配费,存储费两者之和最少?

解 设全年分 n 批供货,每批生产量 $Q=\frac{D}{n}$,周期为$\frac{1}{n}$年$\left(\text{即每隔}\frac{1}{n}\text{年供货一次}\right)$。

每个周期内平均存储量为$\frac{1}{2}Q$

每个周期内的平均存储费用为 $C_1\ \frac{1}{2}Q\ \frac{1}{n}(\text{年})=\frac{C_1Q}{2n}$

全年所需存储费用$\frac{C_1Q}{2n}n=\frac{C_1Q}{2}$

全年所需装配费用 $C_3n=C_3\ \frac{D}{Q}$

全年总费用(以年为单位的平均费用)

$$C(Q)=C_1\frac{Q}{2}+C_3\frac{D}{Q}$$

为求出 $C(Q)$ 的最小值，把 Q 看做连续的变量。

$$\frac{dC(Q)}{dQ}=\frac{C_1}{2}-C_3\frac{D}{Q^2}=0$$

由此得 $$Q_0=\sqrt{\frac{2C_3D}{C_1}}$$

即 $\min C(Q)=C(Q_0)$，Q_0 为经济订购批量。

最佳批次 $n_0=\frac{D}{Q_0}=\sqrt{\frac{C_1D}{2C_3}}$ （取近似的整数）

最佳周期 $t_0=\sqrt{\frac{2C_3}{C_1D}}$

答 全年应分 n_0 次供货可使费用最少。

把例1中的 t_0、Q_0 值与式(14-2)，式(14-3)比较，可知它们是一样的。此处 D 就相当于 R；式(14-1)和式(14-1′)都是研究此类存储问题的数学模型。尽管它们在形式上有差别，但都反映了这类存储问题各量之间的本质联系。

从例1中还看到这些公式在实际应用时还会有一点问题，因为 t_0（或 Q_0，n_0）不一定是整数。假设 $t_0=16.235$（天）。很明显，小数点后面的数字对实际订货间隔的时间是没有意义的，这时可以取近似的整数。取 $t_0\simeq16$ 或 $t_0\simeq17$ 都可以。为了精确起见，可以比较 $C(16)$、$C(17)$ 的大小，再决定 $t_0=16$ 或 $t_0=17$。从图14-4也可以看到 $C(t)$ 在 t_0 附近变化平稳，t 有变化时 $C(t)$ 变化不大。利用数学分析方法可以证明当 t 在 t_0 点有增量 Δt 时，总费用的增量 $\Delta C(t_0)\simeq\frac{C_3}{t_0^3}(\Delta t)^2$。即当 $\Delta t\to0$ 时，ΔC 是 Δt 的高阶无穷小量。（证明的方法可参考微积分台劳公式部分）

例2 某轧钢厂每月按计划需产角钢 30 000 吨，每吨每月需存储费 53 元，每次生产需调整机器设备等，共需装配费 250 000 元。

该厂每月生产角钢一次，生产批量为 30 000 吨。

每月需总费用 $53\times\frac{1}{2}\times30\,000+250\,000=1\,045\,000$（元/月）

全年需费用 $1\,045\,000\times12=12\,540\,000$（元/年）

按 E. O. Q 公式计算每次生产批量

$$\begin{aligned}Q_0&=\sqrt{2\times C_3(\text{装配费})\times D(\text{需求速度})\div C_1(\text{存储费})}\\&=\sqrt{2\times250\,000\times30\,000\div53}\doteq16\,823(\text{吨})\end{aligned}$$

利用 Q_0 计算出全年应生产 n_0 次

$$n_0=\frac{30\,000\times12}{Q_0}\doteq21.4(\text{次})$$

两次生产相隔的时间 $t_0=365/21.4\doteq17$（天）

17 天的单位存储费$\frac{53}{30}\times 17 \doteq 30.00$(元/吨)，一个存储周期共需费用

$$\frac{53}{30}\times 17\times\frac{1}{2}\times 16\,823+250\,000\doteq 502\,562(\text{元})$$

按全年生产 21.5 次(两年生产 43 次)计算，全年共需费用 $502\,562\times 21.5=10\,806\,438$(元/年)。

两者相比较，该厂在利用 E.O.Q 公式求出经济批量进行生产即可每年节约资金

$$12\,540\,000-10\,806\,438=1\,733\,562(\text{元})$$

14.2.2 模型二：不允许缺货，生产需一定时间

本模型的假设条件，除生产需要一定时间的条件外，其余皆与模型一的相同。

设生产批量为 Q，所需生产时间为 T，则生产速度为 $P=Q/T$。

已知需求速度为 R，($R<P$)。生产的产品一部分满足需求，剩余部分才作为存储，这时存储变化如图 14-5 所示。

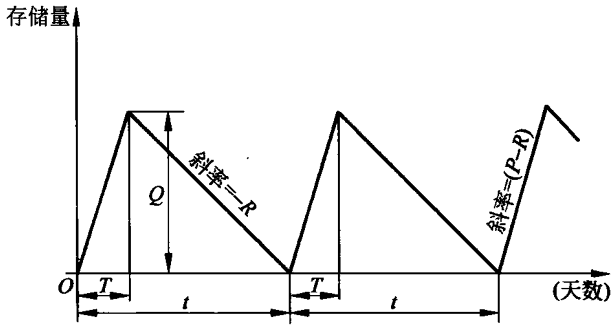


图 14-5

在$[0,T]$区间内，存储以$(P-R)$速度增加，在$[T,t]$区间内存储以速度 R 减少。T 与 t 皆为待定数。从图 14-5 易知$(P-R)T=R(t-T)$，即 $PT=Rt$(等式表示以速度 P 生产 T 时间的产品等于 t 时间内的需求)，并求出 $T=Rt/P$。

t 时间内的平均存储量为$\frac{1}{2}(P-R)T$

t 时间内所需存储费为$\frac{1}{2}C_1(P-R)T$

t 时间内所需装配费为 C_3

单位时间总费用(平均费用)为

$$C(t)=\frac{1}{t}\left[\frac{1}{2}C_1(P-R)Tt+C_3\right]$$

$$=\frac{1}{t}\left[\frac{1}{2}C_1(P-R)\frac{Rt^2}{P}+C_3\right]$$

设 $\min C(t)=C(t_0)$，利用微积分方法可求得

$$t_0=\sqrt{\frac{2C_3P}{C_1R(P-R)}} \tag{14-6}$$

图 14-5 中 t 表示周期，所求出的 t_0 为最佳周期。相应的生产批量

$$Q_0=\text{E. O. Q}=\sqrt{\frac{2C_3RP}{C_1(P-R)}} \tag{14-7}$$

$$\min C(t)=C(t_0)=\sqrt{2C_1C_3R\frac{P-R}{P}} \tag{14-8}$$

利用 t_0 可求出最佳生产时间

$$T_0=\frac{Rt_0}{P}=\sqrt{\frac{2C_3R}{C_1P(P-R)}}$$

将前面求 t_0，Q_0 的公式与式(14-6)，式(14-7)相比较，即知它们只差$\sqrt{\frac{P}{P-R}}$一个因子。当 P 相当大时，$\frac{P}{P-R}$趋近于 1，则两组公式就相同了。

进入存储的最高数量

$$S_0=Q_0-RT_0=\sqrt{\frac{2C_3RP}{C_1(P-R)}}-R\sqrt{\frac{2C_3R}{C_1P(P-R)}}=\sqrt{\frac{2C_3R(P-R)}{C_1P}} \tag{14-9}$$

例 3 某厂每月需甲产品 1 000 件，每月生产率为 5 000 件，每批装配费为 500 元，每月每件产品存储费为 20 元，求 E. O. Q 及最低费用。

解 已知 $C_3=500$，$C_1=20$，$P=5\ 000$，$R=1\ 000$，将各值代入式(14-7)及式(14-8)得

$$\text{E. O. Q}=\sqrt{\frac{2C_3RP}{C_1(P-R)}}=\sqrt{\frac{2\times500\times1\ 000\times5\ 000}{20\times(5\ 000-1\ 000)}}=250(\text{件})$$

$$C_0=\sqrt{\frac{2C_1C_3R(P-R)}{P}}=\sqrt{\frac{2\times20\times500\times1\ 000\times(5\ 000-1\ 000)}{5\ 000}}=\sqrt{16\ 000\ 000}=4\ 000(\text{元})$$

答 每次生产批量为 250 件，每次生产所需装配费及存储费最低为 4 000 元。

例 4 某商店经售甲商品成本单价 500 元，年存储费用为成本的 20%，年需求量 365 件，需求速度为常数。甲商品的订购费为 20 元，提前期为 10 天，求 E. O. Q 及最低费用。

解 此例题从表面上看，似乎应按模型二处理。因为拖后时间似乎与生产需一定时间意义差不多。其实不然，现将本题存储变化情况用图表示之(见图 14-6)，并与模型一、模型二的图相比较，可看到与模型一完全相同。本题只需在存储降至零

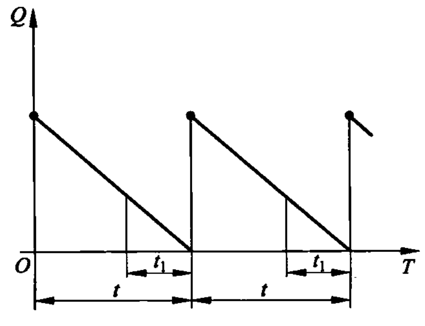


图 14-6

时提前 10 天订货即可保证需求。

利用模型一的 E. O. Q 公式计算：$Q_0=\sqrt{\dfrac{2C_3R}{C_1}}=\sqrt{\dfrac{2\times20\times365}{500\times20\%}}\approx12$(单位)

最低费用：　　$\min C(Q)=C(Q_0)=\sqrt{2C_1C_3R}\approx1\,208$(元)

由于提前期为 10 天，10 天内的需求为 10 单位甲商品，因此只要当存储降至 10 就要订货。一般设 t_1 为提前期，R 为需求速度，当存储降至 $L=Rt_1$ 的时候即订货。L 称为“订购点”(或称订货点)。

从本例题来看，多少时间订一次货，虽可以用 E. O. Q 除以 $R(t_0=Q_0/R)$ 得出 t_0，但求解的过程中并没有求出 t_0，只求出订货点 L，这时存储策略是：不考虑 t_0，只要存储降至 L 即订货，订货量为 Q_0，称这种存储策略为定点订货。相对地称每隔 t_0 时间订货一次为定时订货，每次订货量不变则称为定量订货。

14.2.3　模型三：允许缺货，备货时间很短

模型一、模型二是在不允许缺货的情况下推导出来的。本模型是允许缺货，并把缺货损失定量化来加以研究。由于允许缺货，所以企业可以在存储降至零后，还可以再等一段时间然后订货。这就意味着企业可以少付几次订货的固定费用，少支付一些存储费用。一般地说当顾客遇到缺货时不受损失，或损失很小，而企业除支付少量的缺货费外也无其他损失，这时发生缺货现象可能对企业是有利的。

本模型的假设条件除允许缺货外，其余条件皆与模型一相同。

设　单位时间单位物品存储费用为 C_1，每次订购费为 C_3，缺货费为 C_2(单位缺货损失)，R 为需求速度。求最佳存储策略，使平均总费用最小(见图 14-7)。

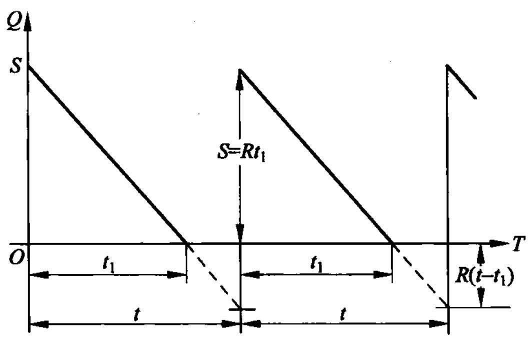


图　14-7

假设最初存储量为 S，可以满足 t_1 时间的需求，t_1 时间的平均存储量为 $\frac{1}{2}S$，在 $(t-t_1)$ 时间的存储为零，平均缺货量为 $\frac{1}{2}R(t-t_1)$。由于 S 仅能满足 t_1 时间的需求 $S=Rt_1$，有

$t_1=S/R$。

在 t 时间内所需存储费 $C_1\frac{1}{2}St_1=\frac{1}{2}C_1\frac{S^2}{R}$

在 t 时间内的缺货费 $C_2\frac{1}{2}R(t-t_1)^2=\frac{1}{2}C_2\frac{(Rt-S)^2}{R}$

订购费为 C_3。

平均总费用 $C(t,S)=\frac{1}{t}\left[C_1\frac{S^2}{2R}+C_2\frac{(Rt-S)^2}{2R}+C_3\right]$

式中有两个变量，利用多元函数求极值的方法求 $C(t,S)$ 的最小值。

由
$$\frac{\partial C}{\partial S}=\frac{1}{t}\left[C_1\frac{S}{R}-C_2\frac{Rt-S}{R}\right]=0$$
$$R\neq 0,t\neq 0,C_1S-C_2(Rt-S)=0$$
得
$$S=\frac{C_2Rt}{C_1+C_2} \tag{14-10}$$
再由
$$\frac{\partial C}{\partial t}=-\frac{1}{t^2}\left[C_1\frac{S^2}{2R}+C_2\frac{(Rt-S)^2}{2R}+C_3\right]+\frac{1}{t}[C_2(Rt-S)]=0$$
因
$$R\neq 0,\quad t\neq 0$$
所以
$$-C_1\frac{S^2}{2}-C_2\frac{(Rt-S)^2}{2}-C_3R+tR[C_2(Rt-S)]=0$$
将式(14-10)中 S 值代入上式，消去 S，得
$$t_0=\sqrt{\frac{2C_3(C_1+C_2)}{C_1RC_2}} \tag{14-11}$$
将式(14-10)代入式(14-11)解出
$$S_0=\sqrt{\frac{2C_2C_3R}{C_1(C_1+C_2)}} \tag{14-12}$$
将式(14-10)，式(14-11)代入 $C(t,S)$ 得
$$\min C(t,S)=C_0(t_0,S_0)=\sqrt{\frac{2C_1C_2C_3R}{C_1+C_2}} \tag{14-13}$$
当 C_2 很大时(不允许缺货)，
$$C_2\to\infty,\frac{C_2}{C_1+C_2}\to 1$$
则
$$t_0\doteq\sqrt{\frac{2C_3}{C_1R}},\quad S_0\doteq\sqrt{\frac{2RC_3}{C_1}},\quad C_0\doteq\sqrt{2C_1C_3R}$$
与式(14-2)，式(14-3)，式(14-5)相同。

模型三中由于允许缺货最佳周期 t_0 为不允许缺货周期 t 的 $\sqrt{\frac{C_2+C_1}{C_2}}$ 倍，又由于 $\frac{C_2+C_1}{C_2}>1$，所以两次订货间隔时间延长了。

在不允许缺货情况下，为满足 t_0 时间内的需求，订货量 $Q_0=Rt_0$

即
$$Q_0=\sqrt{\frac{2RC_3}{C_1}\cdot\frac{C_1+C_2}{C_2}} \tag{14-14}$$

在允许缺货情况下，存储量只需达到 S_0 即可，

$$S_0=\sqrt{\frac{2RC_3C_2}{C_1(C_1+C_2)}}$$

显然 $Q_0>S_0$，它们的差值

$$\begin{aligned}Q_0-S_0&=\sqrt{\frac{2RC_3}{C_1}\cdot\frac{C_1+C_2}{C_2}}-\sqrt{\frac{2RC_3}{C_1}\cdot\frac{C_2}{C_1+C_2}}\\&=\sqrt{\frac{2RC_3}{C_1}}\left(\sqrt{\frac{C_1+C_2}{C_2}}-\sqrt{\frac{C_2}{C_1+C_2}}\right)\\&=\sqrt{\frac{2RC_3}{C_1}}\cdot\frac{C_1}{\sqrt{C_2(C_1+C_2)}}=\sqrt{\frac{2RC_3C_1}{C_2(C_1+C_2)}}\end{aligned}$$

表示在 t_0 时间内的最大缺货量。

在允许缺货条件下，经过研究而得出的存储策略是隔 t_0 时间订货一次，订货量为 Q_0，用 Q_0 中的一部分补足所缺货物，剩余部分 S_0 进入存储。很明显，在相同的时间段落里，允许缺货的订货次数比不允许缺货时订货次数减少了。

例 5 已知需求速度 $R=1\ 000$ 件，$C_1=40$ 元，$C_2=15$ 元，$C_3=500$ 元，求 S_0 及 C_0。

解 利用式(14-12)，式(14-13)即可计算

$$S_0=\sqrt{\frac{2RC_3C_2}{C_1(C_1+C_2)}}=\sqrt{\frac{2\times1\ 000\times15\times500}{40\times(40+15)}}\doteq83(\text{件})$$

$$C_0=\sqrt{\frac{2C_1C_3RC_2}{C_1+C_2}}=\sqrt{\frac{2\times40\times500\times1\ 000\times15}{40+15}}\doteq3\ 302.89(\text{元})$$

答 $S_0\doteq83$ 件，$C_0\doteq3\ 302.89$ 元。

研究模型一、模型二、模型三存储策略之间的差别，可以看到不允许缺货生产需要时间很短条件下得出的存储策略

$$t_0=\sqrt{\frac{2C_3}{C_1R}}\quad(\text{见式(14-2)})$$

$$Q_0=\sqrt{\frac{2C_3R}{C_1}}\quad(\text{见式(14-3)})$$

最大存储量

$$S_0=Q_0$$

在不允许缺货、生产需一定时间条件下，得出存储策略

$$t_0=\sqrt{\frac{2C_3}{C_1R}}\sqrt{\frac{P}{P-R}}\quad(\text{见式(14-6)})$$

$$Q_0=\sqrt{\frac{2C_3R}{C_1}}\sqrt{\frac{P}{P-R}}\quad(\text{见式(14-7)})$$

最大存储量 $S_0=\sqrt{\frac{2C_3R}{C_1}}\sqrt{\frac{P-R}{P}}$ (见式(14-9))

在允许缺货、生产需时间很短条件下，得出存储策略

$$t_0=\sqrt{\frac{2C_3}{C_1R}}\sqrt{\frac{C_1+C_2}{C_2}} \quad (见式(14-11))$$

$$Q_0=\sqrt{\frac{2C_3R}{C_1}}\sqrt{\frac{C_1+C_2}{C_2}} \quad (见式(14-14))$$

最大存储量 $S_0=\sqrt{\frac{2C_3R}{C_1}}\sqrt{\frac{C_2}{C_1+C_2}}$ (见式(14-12))

模型二、模型三只是以模型一的存储策略乘上相应的因子，这样可以便于记忆，再有

模型一中 $\frac{1}{2}Q_0t_0=\frac{1}{2}S_0t_0=\frac{C_3}{C_1}$

模型二中 $\frac{1}{2}S_0t_0=\sqrt{\frac{2C_3R}{C_1}}\sqrt{\frac{P-R}{P}}\sqrt{\frac{2C_3}{C_1R}}\sqrt{\frac{P}{P-R}}=\frac{C_3}{C_1}$

模型三中 $\frac{1}{2}S_0t_0=\sqrt{\frac{2C_3R}{C_1}}\sqrt{\frac{C_2}{C_1+C_2}}\sqrt{\frac{2C_3}{C_1R}}\sqrt{\frac{C_1+C_2}{C_2}}=\frac{C_3}{C_1}$

都是同一个数值，这样就得出它们之间的差别与内在联系。对于允许缺货(缺货需补足)生产需要一定时间条件下的存储策略，我们相信它一定是以模型一的存储策略乘上 $\sqrt{\frac{C_1+C_2}{C_2}}$ 及 $\sqrt{\frac{P}{P-R}}$ 两个因子，它也应有 $\frac{1}{2}S_0t_0=\frac{C_3}{C_1}$，可自行验证。下面做具体介绍。

14.2.4 模型四：允许缺货(需补足缺货)、生产需一定时间

假设条件除允许缺货生产需一定时间外，其余条件皆与模型一相同，其存储变化如图 14-8 所示。

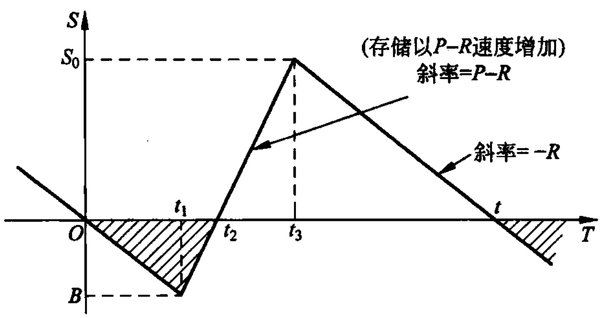


图 14-8

取[0,t]为一个周期，设 t_1 时刻开始生产。

[0,t_2]时间内存储为零，B 表示最大缺货量。

$[t_1,t_2]$时间内除满足需求外，补足$[0,t_1]$时间内的缺货。

$[t_2,t_3]$时间内满足需求后的产品进入存储，存储量以$(P-R)$速度增加。

S 表示存储量，t_3 时刻存储量达到最大，t_3 时刻停止生产。

$[t_3,t]$时间存储量以需求速度 R 减少。

由图 14-8 易知：

最大缺货量 $B=Rt_1$，或 $B=(P-R)(t_2-t_1)$。

即 $Rt_1=(P-R)(t_2-t_1)$，得

$$t_1=\frac{P-R}{P}t_2 \tag{14-15}$$

最大存储量 $S=(P-R)(t_3-t_2)$，或 $S=R(t-t_3)$。

即$(P-R)(t_3-t_2)=R(t-t_3)$，得 $t_3=\frac{R}{P}t+\left(1-\frac{R}{P}\right)t_2$。

或

$$t_3-t_2=\frac{R}{P}(t-t_2) \tag{14-16}$$

在$[0,t]$时间内所需费用

存储费：$$\frac{1}{2}C_1(P-R)(t_3-t_2)(t-t_2)$$

将式(14-16)代入消去 t_3，得$\frac{1}{2}C_1(P-R)\frac{R}{P}(t-t_2)^2$。

缺货费：$$\frac{1}{2}C_2Rt_1t_2$$

将式(14-15)代入消去 t_1，得$\frac{1}{2}C_2R\,\frac{(P-R)}{P}t_2^2$。

装配费：C_3

在$[0,t]$时间内总平均费用为：

$$\begin{aligned}C(t,t_2)&=\frac{1}{t}\left[\frac{1}{2}C_1\,\frac{(P-R)R}{P}(t-t_2)^2+\frac{1}{2}C_2\,\frac{(P-R)R}{P}t_2^2+C_3\right]\\&=\frac{1}{2}\frac{(P-R)R}{P}\left[C_1t-2C_1t_2+(C_1+C_2)\frac{t_2^2}{t}\right]+\frac{C_3}{t}\end{aligned}$$

$$\frac{\partial C(t,t_2)}{\partial t}=\frac{1}{2}\frac{(P-R)R}{P}\left[C_1+(C_1+C_2)t_2^2\left(-\frac{1}{t^2}\right)\right]-\frac{C_3}{t^2} \tag{14-17}$$

$$\frac{\partial C(t,t_2)}{\partial t_2}=\frac{1}{2}\frac{(P-R)R}{P}\left[-2C_1+2(C_1+C_2)t_2\cdot\frac{1}{t}\right] \tag{14-18}$$

令$\frac{\partial C(t,t_2)}{\partial t}=0$，$\frac{\partial C(t,t_2)}{\partial t_2}=0$ 解出 t_1,t_2。

由式(14-18)得

$$t_2=\frac{C_1}{C_1+C_2}t \tag{14-19}$$

由式(14-17)得

$$\frac{1}{2}\frac{(P-R)R}{P}\left[C_1+(C_1+C_2)t_2^2\left(-\frac{1}{t^2}\right)\right]-\frac{C_3}{t^2}=0$$

将式(14-19)代入上式消去 t_2 得

$$\frac{1}{2}\frac{(P-R)R}{P}\left[C_1-(C_1+C_2)\frac{C_1^2}{(C_1+C_2)^2}\right]-C_3\frac{1}{t^2}=0$$

$$\frac{1}{2}\frac{(P-R)R}{P}\left[\frac{C_1^2+C_1C_2-C_1^2}{C_1+C_2}\right]-\frac{C_3}{t^2}=0$$

$$t^2=\frac{2P(C_1+C_2)C_3}{C_1C_2(P-R)R}$$

$$t=\sqrt{\frac{2C_3}{C_1R}}\sqrt{\frac{C_1+C_2}{C_2}}\sqrt{\frac{P}{P-R}}\text{；可记作 } t_0$$

由式(14-19)有

$$t_2=\frac{C_1}{C_1+C_2}t_0=\frac{C_1}{C_1+C_2}\sqrt{\frac{2C_3}{C_1R}}\sqrt{\frac{C_1+C_2}{C_2}}\sqrt{\frac{P}{P-R}}$$

依数学分析的知识可以断定 $C(t,t_2)$ 在 $t=t_0,t_2=\frac{C_1}{C_1+C_2}t_0$ 时有最小值。

相应地得到

$$t_0=\sqrt{\frac{2C_3}{C_1R}}\sqrt{\frac{C_1+C_2}{C_2}}\sqrt{\frac{P}{P-R}} \tag{14-20}$$

$$Q_0=Rt_0=\sqrt{\frac{2C_3R}{C_1}}\sqrt{\frac{C_1+C_2}{C_2}}\sqrt{\frac{P}{P-R}} \tag{14-21}$$

$$\begin{aligned}S_0(\text{最大存储量})&=R(t_0-t_3)=R\left(t_0-\frac{R}{P}t_0-\frac{P-R}{P}t_2\right)\\&=R\left(t_0-\frac{R}{P}t_0-\frac{P-R}{P}\cdot\frac{C_1}{C_1+C_2}t_0\right)\\&=R\frac{P-R}{P}\frac{C_2}{C_1+C_2}t_0\\&=\sqrt{\frac{2C_3R}{C_1}}\sqrt{\frac{C_2}{C_1+C_2}}\sqrt{\frac{P-R}{P}}\end{aligned} \tag{14-22}$$

$$\begin{aligned}B_0(\text{最大缺货量})&=Rt_1=\frac{R(P-R)}{P}t_2\\&=\sqrt{\frac{2C_1C_3R}{(C_1+C_2)C_2}}\sqrt{\frac{P-R}{P}}\end{aligned} \tag{14-23}$$

$$\min C(t_0,t_2)=C_0=\sqrt{2C_1C_3R}\sqrt{\frac{C_2}{C_1+C_2}}\sqrt{\frac{P-R}{P}} \tag{14-24}$$

14.2.5 价格有折扣的存储问题

以上模型所讨论的货物单价是常量，得出的存储策略都与货物单价无关。现在介绍货物单价随订购(或生产)数量而变化时的存储策略。我们常看到一种商品有所谓零售价、批发价和出厂价，购买同一种商品的数量不同，商品单价也不同。一般情况下购买数

量越多，商品单价越低。在少数情况下，某种商品限额供应，超过限额部分的商品单价要提高。

除去货物单价随订购数量而变化外，其余条件皆与模型一的假设相同时，应如何制定相应的存储策略？

记货物单价为 $K(Q)$，设 $K(Q)$ 按三个数量等级变化(见图 14-9)。

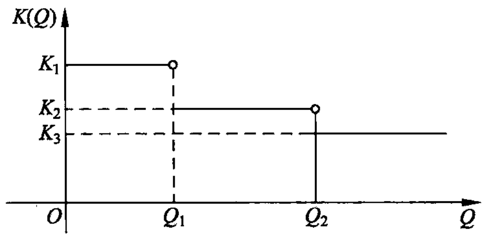


图 14-9

$$K(Q)=\begin{cases}K_1 & 0\leqslant Q<Q_1\\ K_2 & Q_1\leqslant Q<Q_2\\ K_3 & Q_2\leqslant Q\end{cases}$$

当订购量为 Q 时，一个周期内所需费用为

$$\frac{1}{2}C_1Q\frac{Q}{R}+C_3+K(Q)Q$$

$Q\in[0,Q_1)$ 有 $\frac{1}{2}C_1Q\frac{Q}{R}+C_3+K_1Q$

$Q\in[Q_1,Q_2)$ 有 $\frac{1}{2}C_1Q\frac{Q}{R}+C_3+K_2Q$

$Q\geqslant Q_2$ 有 $\frac{1}{2}C_1Q\frac{Q}{R}+C_3+K_3Q$

平均每单位货物所需费用 $C(Q)$(见图 14-10)。

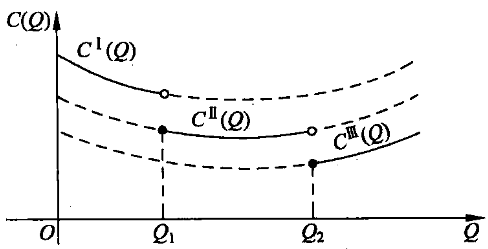


图 14-10

$$C^{\mathrm{I}}(Q)=\frac{1}{2}C_1\frac{Q}{R}+\frac{C_3}{Q}+K_1 \qquad Q\in(0,Q_1)$$

$$C^{\mathrm{II}}(Q)=\frac{1}{2}C_1\frac{Q}{R}+\frac{C_3}{Q}+K_2 \qquad Q\in[Q_1,Q_2)$$

$$C^{\mathrm{III}}(Q)=\frac{1}{2}C_1\frac{Q}{R}+\frac{C_3}{Q}+K_3 \qquad Q\geqslant Q_2$$

如果不考虑 $C^{\mathrm{I}}(Q)$、$C^{\mathrm{II}}(Q)$、$C^{\mathrm{III}}(Q)$的定义域，它们之间只差一个常数，因此它们的导函数相同。为求极小，令导数为零，解得 Q_0，Q_0 落在哪一个区间，事先难以预计。假设 $Q_1<Q_0<Q_2$，这也不能肯定 $C^{\mathrm{II}}(Q_0)$最小。图 14-10 的直观感觉启发我们考虑：是否 $C^{\mathrm{III}}(Q_2)$的费用更小？设最佳订购批量为 Q^*，在给出价格有折扣情况下，求解步骤如下：

(1) 对 $C^{\mathrm{I}}(Q)$(不考虑定义域)求得极值点为 Q_0。

(2) 若 $Q_0<Q_1$，计算

$$C^{\mathrm{I}}(Q_0)=\frac{1}{2}C_1\frac{Q_0}{R}+\frac{C_3}{Q_0}+K_1$$

$$C^{\mathrm{II}}(Q_1)=\frac{1}{2}C_1\frac{Q_1}{R}+\frac{C_3}{Q_1}+K_2$$

$$C^{\mathrm{III}}(Q_2)=\frac{1}{2}C_1\frac{Q_2}{R}+\frac{C_3}{Q_2}+K_3$$

由 $\min\{C^{\mathrm{I}}(Q_0),C^{\mathrm{II}}(Q_1),C^{\mathrm{III}}(Q_2)\}$得到单位货物最小费用的订购批量 Q^*。例如 $\min\{C^{\mathrm{I}}(Q_0),C^{\mathrm{II}}(Q_1),C^{\mathrm{III}}(Q_2)\}=C^{\mathrm{II}}(Q_1)$，则取 $Q^*=Q_1$。

(3) 若 $Q_1\leqslant Q_0<Q_2$，计算 $C^{\mathrm{II}}(Q_0)$、$C^{\mathrm{III}}(Q_2)$。

由 $\min\{C^{\mathrm{II}}(Q_0),C^{\mathrm{III}}(Q_2)\}$决定 Q^*。

(4) 若 $Q_2\leqslant Q_0$，则取 $Q^*=Q_0$。

以上步骤易于推广到单价折扣分 m 个等级的情况。

比如说订购量为 Q，其单价

$$K(Q)=\begin{cases}K_1 & 0\leqslant Q<Q_1\\ K_2 & Q_1\leqslant Q<Q_2\\ \vdots & \\ K_j & Q_{j-1}\leqslant Q<Q_j\\ \vdots & \\ K_m & Q_{m-1}\leqslant Q\end{cases}$$

对应的平均单位货物所需费用为

$$C^j(Q)=\frac{1}{2}C_1\frac{Q}{R}+\frac{C_3}{Q}+K_j \quad j=1,2,\cdots,m$$

对 $C^1(Q)$求得极值点为 Q_0，若 $Q_{j-1}\leqslant Q_0<Q_j$，求 $\min\{C^j(Q_0),C^{j+1}(Q_i),\cdots,C^m(Q_{m-1})\}$，设从此式得到的最小值为 $C^l(Q_{l-1})$，则取 $Q^*=Q_{l-1}$。

例 6 某厂每年需某种元件 30 000 个，每次订购费 $C_3=1\,000$ 元，保管费每件每年 $C_1=100$ 元，不允许缺货。元件单价 K 随采购数量不同而有变化。

$$K(Q)=\begin{cases}200(\text{元}) & Q<1\,500\\ 198(\text{元}) & 1\,500\leqslant Q\end{cases}$$

解 利用 E. O. Q 公式计算

$$Q_0=\sqrt{\frac{2C_3R}{C_1}}=\sqrt{\frac{2\times 1\,000\times 30\,000}{100}}\doteq 775(\text{个})$$

分别计算每次订购 775 个和 1 500 个元件所需平均单位元件所需费用：

$$C(775)=\frac{1}{2}\times 100\times\frac{775}{30\,000}+\frac{1\,000}{775}+200\doteq 202.582(\text{元/个})$$

$$C(1\,500)=\frac{1}{2}\times 100\times\frac{1\,500}{30\,000}+\frac{1\,000}{1\,500}+198\doteq 201.167(\text{元/个})$$

因为 $C(1\,500)<C(775)$，知最佳订购量 $Q=1\,500$(个)。

答 该厂每次应订购 1 500 个元件。

在本章中，由于订购批量不同，订货周期长短不一样，所以才利用平均单位货物所需费用比较优劣。当然也可以利用不同批量，计算其全年所需费用来比较优劣。

也有的折扣条件为

$$K(Q)=\begin{cases}K_1 & \text{当 } Q<Q_1 \text{ 时}\\ K_2 & \text{当 } Q>Q_1 \text{ 时，超过 } Q_1 \text{ 部分}(Q-Q_1)\text{才按 } K_2 \text{ 计算货物单价}\end{cases}$$

如果 $K_2<K_1$，显然是鼓励大量购买货物。在特殊情况下会出现 $K_2>K_1$，这时是利用价格的变化限制购货数量。本章节提供的方法稍加变化后可解决这类问题。

14.3 随机性存储模型

随机性存储模型的重要特点是需求为随机的，其概率或分布为已知。在这种情况下，前面所介绍过的模型已经不能适用了。例如商店对某种商品进货 500 件，这 500 件商品可能在一个月内售完，也有可能在两个月之后还有剩余。商店如果想既不因缺货而失去销售机会，又不因滞销而过多积压资金，这时必须采用新的存储策略。可供选择的策略主要有三种：

(1) 定期订货，但订货数量需要根据上一个周期末剩下货物的数量决定订货量。剩下的数量少，可以多订货。剩下的数量多，可以少订或不订货。这种策略可称为定期订货法。

(2) 定点订货，存储降到某一确定的数量时即订货，不再考虑间隔的时间。这一数量值称为订货点，每次订货的数量不变，这种策略可称之为定点订货法。

(3) 把定期订货与定点订货综合起来的方法，隔一定时间检查一次存储，如果存储数

量高于一个数值 s，则不订货。小于 s 时则订货补充存储，订货量要使存储量达到 S，这种策略可以简称为 (s,S) 存储策略。

此外与确定性模型不同的特点还有：不允许缺货的条件只能从概率的意义方面理解，如不缺货的概率为 0.9 等。存储策略的优劣通常以赢利的期望值的大小作为衡量的标准。

14.3.1 模型五：需求是随机离散的

报童问题：报童每日售报数量是一个随机变量。报童每售出一份报纸赚 k 元。如报纸未能售出，每份赔 h 元。每日售出报纸份数 r 的概率 $P(r)$ 根据以往的经验是已知的，问报童每日最好准备多少份报纸？

这个问题是报童每日报纸的订货量 Q 为何值时，赚钱的期望值最大？反言之，如何适当地选择 Q 值，使因不能售出报纸的损失及因缺货失去销售机会的损失，两者期望值之和最小。现在用计算损失期望值最小的办法求解。

解 设售出报纸数量为 r，其概率 $P(r)$ 为已知，$\sum_{r=0}^{\infty} P(r) = 1$。

设报童订购报纸数量为 Q。

① 供过于求时 $(r \leqslant Q)$，这时报纸因不能售出而承担的损失，其期望值为：

$$\sum_{r=0}^{Q} h(Q-r)P(r) \qquad \text{（初学者可参看例 7）}$$

② 供不应求时 $(r > Q)$，这时因缺货而少赚钱的损失，其期望值为

$$\sum_{r=Q+1} k(r-Q)P(r)$$

综合①，②两种情况，当订货量为 Q 时，损失的期望值为

$$C(Q) = h\sum_{r=0}^{Q}(Q-r)P(r) + k\sum_{r=Q+1}^{\infty}(r-Q)P(r)$$

要从式中决定 Q 的值，使 $C(Q)$ 最小。

由于报童订购报纸的份数只能取整数，r 是离散变量，所以不能用求导数的方法求极值。为此设报童每日订购报纸份数最佳量为 Q，其损失期望值应有

① $C(Q) \leqslant C(Q+1)$

② $C(Q) \leqslant C(Q-1)$

从①出发进行推导有

$$h\sum_{r=0}^{Q}(Q-r)P(r) + k\sum_{r=Q+1}^{\infty}(r-Q)P(r) \leqslant$$
$$h\sum_{r=0}^{Q+1}(Q+1-r)P(r) + k\sum_{r=Q+2}^{\infty}(r-Q-1)P(r)$$

经化简后得

$$(k+h)\sum_{r=0}^{Q}P(r)-k\geqslant 0$$

即

$$\sum_{r=0}^{Q}P(r)\geqslant\frac{k}{k+h}$$

从②出发进行推导有

$$h\sum_{r=0}^{Q}(Q-r)P(r)+k\sum_{r=Q+1}^{\infty}(r-Q)P(r)\leqslant$$
$$h\sum_{r=0}^{Q-1}(Q-1-r)P(r)+k\sum_{r=Q}^{\infty}(r-Q+1)P(r)$$

经化简后得

$$(k+h)\sum_{r=0}^{Q-1}P(r)-k\leqslant 0$$

即

$$\sum_{r=0}^{Q-1}P(r)\leqslant\frac{k}{k+h}$$

报童应准备的报纸最佳数量 Q 应按下列不等式确定

$$\sum_{r=0}^{Q-1}P(r)<\frac{k}{k+h}\leqslant\sum_{r=0}^{Q}P(r) \tag{14-25}$$

从赢利最大来考虑报童应准备的报纸数量。设报童订购报纸数量为 Q，获利的期望值为 $C(Q)$，其余符号和前面推导时表示的意义相同。

当需求 $r\leqslant Q$ 时，报童只能售出 r 份报纸，每份赚 k(元)，共赚 kr(元)。未售出的报纸，每份赔 h(元)，滞销损失为 $h(Q-r)$(元)。

此时赢利的期望值为

$$\sum_{r=0}^{Q}[kr-h(Q-r)]P(r)$$

当需求 $r>Q$ 时，报童因为只有 Q 份报纸可供销售，赢利的期望值为 $\sum\limits_{r=Q+1}^{\infty}k\cdot QP(r)$，无滞销损失。由以上分析知赢利的期望值

$$C(Q)=\sum_{r=0}^{Q}k\cdot r\cdot P(r)-\sum_{r=0}^{Q}h(Q-r)P(r)+\sum_{r=Q+1}^{\infty}k\cdot QP(r)$$

为使订购 Q 赢利的期望值最大，应满足下列关系式：

① $C(Q+1)\leqslant C(Q)$

② $C(Q-1)\leqslant C(Q)$

从①式推导

$$k\sum_{r=0}^{Q+1}rP(r)-h\sum_{r=0}^{Q+1}(Q+1-r)P(r)+k\sum_{r=Q+2}^{\infty}(Q+1)P(r)\leqslant$$

$$k\sum_{r=0}^{Q}rP(r)-h\sum_{r=0}^{Q}(Q-r)P(r)+k\sum_{r=Q+1}^{\infty}Q\cdot P(r)$$

经化简后得

$$kP(Q+1)-h\sum_{r=0}^{Q}P(r)+k\sum_{r=Q+2}^{\infty}P(r)\leqslant 0$$

进一步化简得

$$k\left[1-\sum_{r=0}^{Q}P(r)\right]-h\sum_{r=0}^{Q}P(r)\leqslant 0$$

$$\sum_{r=0}^{Q}P(r)\geqslant\frac{k}{k+h}$$

同理从②推导出

$$\sum_{r=0}^{Q-1}P(r)\leqslant\frac{k}{k+h}$$

用以下不等式确定 Q 的值,这一公式与式(14-25)完全相同。

$$\sum_{r=0}^{Q-1}P(r)<\frac{k}{k+h}\leqslant\sum_{r=0}^{Q}P(r)$$

尽管报童问题中损失最小的期望值与赢利最大的期望值是不同的,但确定 Q 值的条件是相同的。无论从哪一个方面来考虑,报童的最佳订购份数是一个确定的数值。在下面的模型六中将进一步说明这个问题。

例 7 某店拟出售甲商品,每单位甲商品成本 200 元,售价 280 元。如不能售出必须减价为 160 元,减价后一定可以售出。已知售货量 r 的概率服从泊松分布

$$P(r)=\frac{e^{-\lambda}\lambda^{r}}{r!}\quad(\lambda\text{ 为平均售出数})$$

根据以往经验,平均售出数为 6 单位($\lambda=6$)。问该店订购量应为多少单位?

解 该店的缺货损失,每单位商品为 280－200＝80(元)。滞销损失,每单位商品为 200－160＝40(元),利用式(14-25),其中 $k=80,h=40$。

$$\frac{k}{h+k}=\frac{80}{80+40}\doteq 0.667$$

$$P(r)=\frac{e^{-6}6^{r}}{r!},\quad \sum_{r=0}^{Q}P(r)\text{ 记作 }F(Q)\text{,可查统计表}$$

$$F(6)=\sum_{r=0}^{5}\frac{e^{-6}6^{r}}{r!}=0.6063,F(7)=\sum_{r=0}^{7}\frac{e^{-6}6^{r}}{r!}=0.7440$$

因 $F(6)<\frac{k}{k+h}<F(7)$,故订货量应为 7 单位,此时损失的期望值最小。

答 该店订货量应为 7 单位甲商品。

例 8 上题中如缺货损失为 40 元,滞销损失为 80 元。在这种情况下该店订货量应为多少?

解 利用式(14-25),其中 $k=40,h=80$。

$$\frac{k}{k+h}=\frac{40}{120}\doteq 0.3333$$

查统计表,找与 0.333 3 相近的数。

$$F(4)=\sum_{r=0}^{4}\frac{e^{-6}6^r}{r!}=0.2851,F(5)=\sum_{r=0}^{5}\frac{e^{-6}6^r}{r!}=0.4457$$

$F(4)<0.3333<F(5)$,故订货量应为甲商品 5 个单位。

答 该店订货量为 5 个单位甲商品。

模型五只解决一次订货问题,对报童问题实际上每日订货策略问题也应认为解决了。但模型中有一个严格的约定,即两次订货之间没有联系,都看做独立的一次订货。这种存储策略也可称之为定期定量订货。

14.3.2 模型六:需求是连续的随机变量

设 货物单位成本为 K,货物单位售价为 P,单位存储费为 C_1,需求 r 是连续的随机变量,密度函数为 $\phi(r)$,$\phi(r)\mathrm{d}r$ 表示随机变量在 r 与 $r+\mathrm{d}r$ 之间的概率,其分布函数 $F(a)=\int_0^a\phi(r)\mathrm{d}r,(a>0)$,生产或订购的数量为 Q,问如何确定 Q 的数值,使赢利的期望值最大?

解 首先我们来考虑当订购数量为 Q 时,实际销售量应该是 $\min[r,Q]$。也就是当需求为 r 而 r 小于 Q 时,实际销售量为 r;$r\geqslant Q$ 时,实际销售量只能是 Q。

$$\text{需支付的存储费用 } C_1(Q)=\begin{cases}C_1(Q-r) & r\leqslant Q\\ 0 & r>Q\end{cases}$$

货物的成本为 KQ,本阶段订购量为 Q 赢利为 $W(Q)$,赢利的期望值记作 $E[W(Q)]$。

本阶段的赢利

$$W(Q)=P\cdot\min[r,Q]-KQ-C_1(Q)$$

赢利=实际销售货物的收入-货物成本-支付的存储费用

赢利的期望值

$$E[W(Q)]=\int_0^Q Pr\phi(r)\mathrm{d}r+\int_Q^\infty PQ\phi(r)\mathrm{d}r-KQ-\int_0^Q C_1(Q-r)\phi(r)\mathrm{d}r$$

$$=\int_0^\infty Pr\phi(r)\mathrm{d}r-\int_Q^\infty Pr\phi(r)\mathrm{d}r+\int_Q^\infty PQ\phi(r)\mathrm{d}r-KQ-\int_0^Q C_1(Q-r)\phi(r)\mathrm{d}r$$

$$=\underbrace{PE(r)}_{\text{常量(称为平均赢利)}}-\left\{\underbrace{P\int_Q^\infty(r-Q)\phi(r)\mathrm{d}r}_{\text{因缺货失去销售机会损失的期望值}}+\underbrace{\int_0^Q C_1(Q-r)\phi(r)\mathrm{d}r}_{\text{因滞销受到损失的期望值(只考虑了存储费)}}+\underbrace{KQ}_{\text{常量}}\right\}$$

记 $$E[C(Q)] = P\int_Q^\infty (r-Q)\phi(r)\mathrm{d}r + C_1\int_0^Q (Q-r)\phi(r)\mathrm{d}r + KQ$$

为使赢利期望值极大化，有下列等式

$$\max E[W(Q)] = PE(r) - \min E[C(Q)] \tag{14-26}$$

$$\max E[W(Q)] + \min E[C(Q)] = PE(r) \tag{14-27}$$

式(14-26)表明了赢利最大与损失极小所得出的 Q 值相同。式(14-27)表明最大赢利期望值与损失极小期望值之和是常数。

根据上面的分析，求赢利极大可以转化为求 $E[C(Q)]$(损失期望值)极小。当 Q 可以连续取值时，$E[C(Q)]$是 Q 的连续函数。可利用微分法求最小。

$$\begin{aligned}\frac{\mathrm{d}E[C(Q)]}{\mathrm{d}Q} &= \frac{\mathrm{d}}{\mathrm{d}Q}\left[P\int_Q^\infty (r-Q)\phi(r)\mathrm{d}r + \int_0^Q C_1(Q-r)\phi(r)\mathrm{d}r + KQ\right] \\ &= C_1\int_0^Q \phi(r)\mathrm{d}r - P\int_Q^\infty \phi(r)\mathrm{d}r + K\end{aligned}$$

令 $$\frac{\mathrm{d}E[C(Q)]}{\mathrm{d}Q} = 0 \quad 记 \quad F(Q) = \int_0^Q \phi(r)\mathrm{d}r$$

即 $$C_1F(Q) - P[1-F(Q)] + K = 0$$

$$F(Q) = \frac{P-K}{C_1+P}$$

从此式中解出 Q，记为 Q^*，Q^* 为 $E[C(Q)]$的驻点。又因

$$\frac{\mathrm{d}^2E[C(Q)]}{\mathrm{d}Q^2} = C_1\phi(Q) + P\phi(Q) > 0$$

知 Q^* 为 $E[C(Q)]$的极小值点，在本模型中也是最小值点。

若 $P-K\leqslant 0$，显然由于 $F(Q)\geqslant 0$，等式不成立，此时 Q^* 取零值。即售价低于成本时，不需要订货(或生产)。

式中只考虑了失去销售机会的损失，如果缺货时要付出的费用 $C_2 > P$，应有

$$E[C(Q)] = C_2\int_Q^\infty (r-Q)\phi(r)\mathrm{d}r + C_1\int_0^Q (Q-r)\phi(r)\mathrm{d}r + KQ \tag{14-28}$$

按上述办法推导得

$$F(Q) = \int_0^Q \phi(r)\mathrm{d}r = \frac{C_2-K}{C_1+C_2}$$

模型五及模型六都是只解决一个阶段的问题。从一般情况来考虑，上一阶段未售出的货物可以在第二阶段继续出售。这时应该如何制定存储策略呢？

假设上一阶段未能售出的货物数量为 I，作为本阶段初的存储，有

$$\begin{aligned}\min E[C(Q)] &= K(Q-I) + C_2\int_Q^\infty (r-Q)\phi(r)\mathrm{d}r + C_1\int_0^Q (Q-r)\phi(r)\mathrm{d}r \\ &= \underbrace{-KI}_{常量} + \min\Big\{\underbrace{C_2\int_Q^\infty (r-Q)\phi(r)\mathrm{d}r + C_1\int_0^Q (Q-r)\phi(r)\mathrm{d}r + KQ}_{与(14\text{-}28)式相同}\Big\}\end{aligned}$$

利用 $F(Q)=\int_0^Q \phi(r)\mathrm{d}r=\dfrac{C_2-K}{C_1+C_2}$ 求出 Q^* 值，相应的存储策略为：当 $I\geqslant Q^*$ 时，本阶段不订货。当 $I<Q^*$ 时，本阶段应订货，订货量为 $Q=Q^*-I$，使本阶段的存储达到 Q^*，这时赢利期望值最大。

这种策略也可以称作定期订货，订货量不定的存储策略。

14.3.3 模型七：(s,S) 型存储策略

1. 需求为连续的随机变量时

设 货物单位成本为 K，单位存储费为 C_1，单位缺货费为 C_2，每次订购费为 C_3，需求 r 是连续的随机变量，密度函数为 $\phi(r)$，$\int_0^\infty \phi(r)\mathrm{d}r=1$，分布函数 $F(a)=\int_0^a \phi(r)\mathrm{d}r$，$(a>0)$，期初存储为 I，订货量为 Q，此时期初存储达到 $S=I+Q$。问如何确定 Q 的值，使损失的期望值最小(赢利的期望值最大)?

解 期初存储 I 在本阶段中为常量，订货量为 Q，则期初存储达到 $S=I+Q$。本阶段需订货费 C_3+KQ，本阶段需付存储费用的期望值为

$$\int_0^{I+Q=S} C_1(S-r)\phi(r)\mathrm{d}r$$

需付缺货费用的期望值为

$$\int_{S=I+Q}^{\infty} C_2(r-S)\phi(r)\mathrm{d}r$$

本阶段所需订货费及存储费、缺货费期望值之和

$$\begin{aligned}C(I+Q)=C(S)&=C_3+KQ+\int_0^S C_1(S-r)\phi(r)\mathrm{d}r+\int_S^\infty C_2(r-S)\phi(r)\mathrm{d}r\\&=C_3+K(S-I)+\int_0^S C_1(S-r)\phi(r)\mathrm{d}r+\int_S^\infty C_2(r-S)\phi(r)\mathrm{d}r\end{aligned}$$

Q 可以连续取值，$C(S)$ 是 S 的连续函数。

$$\frac{\mathrm{d}C(S)}{\mathrm{d}S}=K+C_1\int_0^S \phi(r)\mathrm{d}r-C_2\int_S^\infty \phi(r)\mathrm{d}r$$

令 $\dfrac{\mathrm{d}C(S)}{\mathrm{d}S}=0$，有

$$F(S)=\int_0^S \phi(r)\mathrm{d}r=\frac{C_2-K}{C_1+C_2} \tag{14-29}$$

$\dfrac{C_2-K}{C_1+C_2}$ 严格小于1，称为临界值，以 N 表示：$\dfrac{C_2-K}{C_1+C_2}=N$。

为得出本阶段的存储策略：

由 $\int_0^S \phi(r)\mathrm{d}r=N$，确定 S 的值。

订货量 $Q=S-I$。

本模型中有订购费 C_3，如果本阶段不订货可以节省订购费 C_3，因此我们设想是否存在一个数值 $s(s\leqslant S)$ 使下面不等式能成立。

$$Ks+C_1\int_0^s(s-r)\phi(r)\mathrm{d}r+C_2\int_s^\infty(r-s)\phi(r)\mathrm{d}r\leqslant$$

$$C_3+KS+C_1\int_0^S(S-r)\phi(r)\mathrm{d}r+C_2\int_S^\infty(r-S)\phi(r)\mathrm{d}r$$

当 $s=S$ 时，不等式显然成立。

当 $s<S$ 时，不等式右端存储费用期望值大于左端存储费用期望值，右端缺货费用期望值小于左端缺货费用期望值；一增一减后仍然使不等式成立的可能性是存在的。如有不止一个 s 的值使下列不等式成立，则选其中最小者作为本模型 (s,S) 存储策略的 s。

$$C_3+K(S-s)+C_1\left[\int_0^S(S-r)\phi(r)\mathrm{d}r-\int_0^s(s-r)\phi(r)\mathrm{d}r\right]+$$

$$C_2\left[\int_S^\infty(r-S)\phi(r)\mathrm{d}r-\int_s^\infty(r-s)\phi(r)\mathrm{d}r\right]\geqslant 0$$

相应的存储策略是：每阶段初期检查存储，当库存 $I<s$ 时，需订货，订货的数量为 Q，$Q=S-I$。当库存 $I\geqslant s$ 时，本阶段不订货。这种存储策略是：定期订货但订货量不确定。订货数量的多少视期末库存 I 来决定订货量 Q，$Q=S-I$。对于不易清点数量的存储，人们常把存储分两堆存放，一堆的数量为 s，其余的另放一堆。平时从另放的一堆中取用，当动用了数量为 s 的一堆时，期末即订货。如果未动用 s 的一堆时，期末即可不订货，俗称两堆法。

2. 需求是离散的随机变量时

设　需求 r 取值为 $r_0,r_1,\cdots,r_m(r_i<r_{i+1})$，其概率为 $P(r_0),P(r_1),\cdots,P(r_m)$，$\sum_{i=0}^{m}P(r_i)=1$，原有存储量为 I(在本阶段内为常量)。

当本阶段开始时订货量为 Q，存储量达到 $I+Q$，本阶段所需的各种费用：

订货费：C_3+KQ。

存储费：当需求 $r<I+Q$ 时，未能售出的存储部分需付存储费。

当需求 $r\geqslant I+Q$ 时，不需要付存储费。

所需存储费的期望值

$$\sum_{r\leqslant I+Q}C_1(I+Q-r)P(r)\qquad (r=I+Q\text{ 时，不付存储费及缺货费})$$

缺货费：当需求 $r>I+Q$ 时，$(r-I-Q)$ 部分需付缺货费。

缺货费用的期望值

$$\sum_{r>I+Q}C_2(r-I-Q)P(r)$$

本阶段所需订货费及存储费、缺货费期望之和

$$C(I+Q)=C_3+KQ+\sum_{r\leqslant I+Q}C_1(I+Q-r)P(r)+\sum_{r>I+Q}C_2(r-I-Q)P(r)$$

$I+Q$ 表示存储所达到的水平，记 $S=I+Q$，上式可写为

$$C(S)=C_3+K(S-I)+\sum_{r\leqslant S}(S-r)C_1P(r)+\sum_{r>S}C_2(r-S)P(r)$$

求出 S 值使 $C(S)$ 最小。

解法：

(1) 将需求 r 的随机值按大小顺序排列为

$$r_0,r_1,\cdots,r_i,r_{i+1},\cdots,r_m,r_i<r_{i+1},r_{i+1}-r_i=\Delta r_i\neq 0(i=0,1,\cdots,m-1)$$

(2) S 只从 $r_0,r_1,\cdots,r_m$ 中取值。当 S 取值为 r_i 时，记为 S_i。

$$\Delta S_i=S_{i+1}-S_i=r_{i+1}-r_i=\Delta r_i\neq 0(i=0,1,\cdots,m-1)$$

(3) 求 S 的值使 $C(S)$ 最小。因为

$$C(S_{i+1})=C_3+K(S_{i+1}-I)+\sum_{r\leqslant S_{i+1}}C_1(S_{i+1}-r)P(r)+\sum_{r>S_{i+1}}C_2(r-S_{i+1})P(r)$$

$$C(S_i)=C_3+K(S_i-I)+\sum_{r\leqslant S_i}C_1(S_i-r)P(r)+\sum_{r>S_i}C_2(r-S_i)P(r)$$

$$C(S_{i-1})=C_3+K(S_{i-1}-I)+\sum_{r\leqslant S_{i-1}}C_1(S_{i-1}-r)P(r)+\sum_{r>S_{i-1}}C_2(r-S_{i-1})P(r)$$

为选出使 $C(S_i)$ 最小的 S 值，S_i 应满足下列不等式

① $C(S_{i+1})-C(S_i)\geqslant 0$

② $C(S_i)-C(S_{i-1})\leqslant 0$

定义 $\Delta C(S_i)=C(S_{i+1})-C(S_i)$，$\Delta C(S_{i-1})=C(S_i)-C(S_{i-1})$。

由①可推导出

$$\begin{aligned}\Delta C(S_i)&=K\Delta S_i+C_1\Delta S_i\sum_{r\leqslant S_i}P(r)-C_2\Delta S_i\sum_{r>S_i}P(r)\\&=K\Delta S_i+C_1\Delta S_i\sum_{r\leqslant S_i}P(r)-C_2\Delta S_i\left[1-\sum_{r\leqslant S_i}P(r)\right]\\&=K\Delta S_i+(C_1+C_2)\Delta S_i\sum_{r\leqslant S_i}P(r)-C_2\Delta S_i\geqslant 0\end{aligned}$$

因 $\Delta S_i\neq 0$ 即

$$K+(C_1+C_2)\sum_{r\leqslant S_i}P(r)-C_2\geqslant 0$$

有

$$\sum_{r\leqslant S_i}P(r)\geqslant\frac{C_2-K}{C_1+C_2}=N$$

由②同理可推导出

$$\sum_{r\leqslant S_{i-1}}P(r)\leqslant\frac{C_2-K}{C_1+C_2}=N$$

综合以上两式，得到为确定 S_i 的不等式

$$\sum_{r\leqslant S_{i-1}}P(r)<N=\frac{C_2-K}{C_1+C_2}\leqslant\sum_{r\leqslant S_i}P(r)\qquad(14\text{-}30)$$

取满足式(14-30)的 S_i 为 S，本阶段订货量为 $Q=S-I$。

例 9　设某公司利用塑料作原料制成产品出售，已知每箱塑料购价为 1 600 元，订购费 $C_3=300$ 元，存储费每箱 $C_1=80$ 元，缺货费每箱 $C_2=2\ 100$ 元，原有存储量 $I=10$ 箱。已知对原料需求的概率

$$P(r=60\text{ 箱})=0.20,\quad P(r=80\text{ 箱})=0.20$$
$$P(r=100\text{ 箱})=0.40,\quad P(r=120\text{ 箱})=0.20$$

求该公司订购原料的最佳订购量。

解　① 计算临界值 $N=\dfrac{2\ 100-1\ 600}{2\ 100+80}\doteq 0.229$

② 选使不等式 $\sum\limits_{r\leqslant S_i}P(r)\geqslant N$ 成立的 S_i 最小值作为 S。

$$P(60)=0.20\ngtr 0.229$$
$$P(60)+P(80)=0.20+0.20=0.40>0.229$$
$$S_i=80,\text{作为 } S$$

③ 原存储 $I=20$，订货量 $Q=S-I=80-20=60$。

答　该公司应订购塑料 60 箱。

下面对答案进行验证，分别计算 S 为 60，80，100 所需订货费及存储费期望值、缺货费期望值三者之和。比较它们看是否当 S 为 80 时最小(见表 14-1)。

表　14-1

S	I	$Q=S-I$	订货费 C_3+KQ	存储费期望值 $C_1\sum\limits_{r\leqslant S}(S-r)P(r)$	缺货费期望值 $C_2\sum\limits_{r>S}(S-r)P(r)$	总计
60	20	40	64 300	0	62 700	131 500
80	20	60	96 300	320	33 600	130 220*
100	20	80	128 300	960	8 400	137 660

* 为最优

比较后知 $S=80$ 所需总费用最少，订购量 $Q=60$。

本模型还有另一方面的问题，原存储量 I 达到什么水平可以不订货？假设这一水平是 s，当 $I>s$ 时可以不订货，当 $I\leqslant s$ 时要订货，使存储达到 S，订货量 $Q=S-I$。

计算 s 的方法：考查不等式

$$Ks+\sum_{r\leqslant s}C_1(s-r)P(r)+\sum_{r>s}C_2(r-s)P(r)\leqslant C_3+KS+$$
$$\sum_{r\leqslant s}C_1(S-r)P(r)+\sum_{r>S}C_2(r-S)P(r)\qquad(14\text{-}31)$$

因 S 也只从 $r_0,r_1,\cdots,r_m$ 中取值。使式(14-31)成立的 $r_i(r_i\leqslant S)$ 的值中最小者定为 s。当 $s<S$ 时，式(14-31)左端缺货费用的期望值虽然会增加，但订货费及存储费用期望值都减少，一增一减之间，使不等式仍有可能成立。在最不利的情况下 $s=S$ 时不等式是成立的(因为 $0<C_3$)。因此我们相信一定能找到 s 值。当然计算 s 要比计算 S 复杂一些，但就具

体问题计算 s 也不是很难。如例 9 中要计算 s 确很简单。由于已算出 $S=80$,可以作为 s 的 r 值只有 60 或 80 两个值。

将 60 作为 s 值代入式(14-31)左端得

$$1\,600\times60+2\,100\times[(80-60)\times0.2+(100-60)\times0.4+(120-60)\times0.2]$$
$$=163\,200$$

将 80 代入(14-31)式右端得

$$300+1\,600\times80+80\times[(80-60)\times0.2]+2\,100\times[(100-80)\times0.4+(120-80)\times0.2]=162\,220$$

即左端数值为 163 200,右端数值为 162 220,不等式成立,60 已是 r 的最小值故 $s=60$。

例 9 的存储策略为每个阶段开始时检查存储量 I,当 $I>60$ 箱时不必补充存储。当 $I\leqslant60$ 箱时补充存储量达到 80 箱。

例 10 某厂对原料需求量的概率为

$$P(r=160)=0.1,\quad P(r=170)=0.2,\quad P(r=180)=0.3$$
$$P(r=190)=0.3,\quad P(r=200)=0.1$$

订货费 $C_3=6\,000$ 元,$K=1\,500$ 元

存储费 $C_1=100$ 元(在本阶段的费用)

缺货费 $C_2=2\,200$ 元(在本阶段的费用)

求该厂存储策略。

解 (1) 利用式(14-30)计算临界值 $N=\dfrac{2\,200-1\,500}{2\,200+100}=0.304$

(2) 求 S

$$P(r=160)+P(r=170)=0.3\ngtr0.304$$
$$P(r=160)+P(r=170)+P(r=180)=0.6>0.304$$

可知 $S=180$。

(3) 利用式(14-31)计算 s

$S=180$,(14-31)式右端为

$$6\,000+1\,500\times180+100\times[(180-60)\times0.1+(180-170)\times0.2+(180-180)\times0.3]+2\,200\times[(190-180)\times0.3+(200-180)\times0.1]$$
$$=287\,400$$

$s=160$,式(14-31)左端为

$$1\,500\times160+100\times(160-160)\times0.1+2\,200\times[(170-160)\times0.2+(180-160)\times0.3+(190-160)\times0.3+(200-160)\times0.1]$$
$$=286\,200$$

由于 $286\,200<287\,400$,故知 $s=160$。

答 该厂存储策略每当存储 $I\leqslant160$ 时补充存储使存储量达到 180,每当存储 $I>160$

时不补充。

例 11 某市石油公司，下设几个售油站。石油存放在郊区大型油库里，需要时用汽车将油送至各售油站。该公司希望确定一种补充存储的策略，以确定应储存的油量。该公司经营石油品种较多，其中销售量较多的一种是柴油。因之希望先确定柴油的存储策略。经调查后知每月柴油出售量服从指数分布，平均销售量每月为一百万升。其密度为：

$$f(r)=\begin{cases}0.000\,001e^{-0.000\,001\times r} & r\geqslant 0\\ 0 & r<0\end{cases}$$

柴油每升 6 元，不需订购费。由于油库归该公司管辖，油池灌满与未灌满时的管理费用实际上没有多少差别，故可以认为存储费用为零。如缺货就从邻市调用，缺货费9 元/升。求柴油的存储策略。

解 根据例 11 中条件知 $C_1=0, C_3=0, K=2, C_2=3$，计算临界值。

$N=\dfrac{9-6}{9+0}=0.333$ 由于密度是连续函数，所以需利用积分计算求出 S。

$$\int_0^S f(r)\mathrm{d}r=\int_0^S 0.000\,001e^{-0.000\,001\times r}\mathrm{d}r=0.333$$

$$-e^{-0.000\,001\times r}\Big|_0^S=0.333$$

$$e^{-0.000\,001\times s}=0.667$$

两端取对数解出

$$S=405\,000(\text{升})$$

利用式(14-31)求 s 只需把相应的求和部分利用积分计算即可。

$$2\times s+\int_0^S 0\times(s-r)f(r)\mathrm{d}r+3\int_s^\infty(r-s)f(r)\mathrm{d}r\leqslant 2\times S+$$

$$\int_0^S 0\times(S-r)f(r)\mathrm{d}r+3\int_S^\infty(r-S)f(r)\mathrm{d}r$$

即

$$2\times s+\int_s^\infty 3\times(r-s)f(r)\mathrm{d}r\leqslant 2\times S+\int_S^\infty 3\times(r-S)f(r)\mathrm{d}r$$

由观察，它有唯一解 $s=S$，所以当库存柴油下降到 405 000 升以下时就应订购，使库存达到 405 000 升。为什么会出现 $s=S$ 呢？原因在于订购费为零，可以频繁订货。又存储费 C_1 也为零，存储量多一些也不会增加费用。

由于 r 连续，$f(r)$也连续，所以计算 S 值时要利用等式 $\int_0^S f(r)\mathrm{d}r=N$ 求解 S 的值。

14.3.4 模型八：需求和备货时间都是随机离散的（仅通过具体例题介绍求解法）

若 t 时间内的需求量 r 是随机的，其概率 $\varphi_t(r)$已知，单位时间内的平均需求为 ρ 也是已知的，则 t 时间内的平均需求为 ρt。备货时间 x 是随机的，其概率 $P(x)$已知。

设 单位货物年存储费用为 C_1，每阶段单位货物缺货费用为 C_2，每次订购费用为 C_3，年平均需求为 D。由于需求、备货时间都是随机的，应有缓冲(安全)存储量 B，以减少发生缺货现象。L：订货点，B：缓冲存储量，$x_1,x_2,\cdots$为备货时间(见图 14-11)。

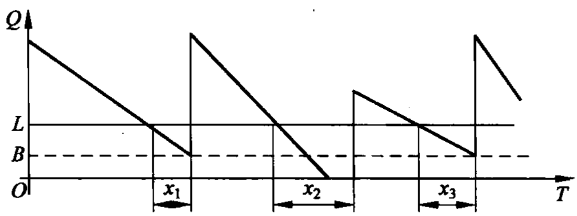


图 14-11

问如何确定缓冲存储量 B，订货点 L，以及订货量 Q_0，使总费用最小？

对这种类型问题的解法：先按确定性模型求出 E.O.Q 及最佳批次 n_0。

$$Q_0=\text{E.O.Q}=\sqrt{\frac{2C_3D}{C_1}};\quad n_0=\frac{D}{Q_0}$$

(存储策略全年分 n_0 次订货，每次订货量 Q_0)

订货点 L 的确定方法除应满足备货时间内的平均需求 D_L 还要求维持缓冲存储量 B，由于备货时间是随机的，设平均备货时间为 μ。

则

$$L=D_L+B=\mu\rho+B$$

当存储量降至 L 时订货，由于备货时间延长，或因需求增加而引起缺货的概率记作

$$P_L=\sum_x P(x)F_x(L)$$

其中：$P(x)$表示备货时间为 x 天的概率；

$F_x(L)$表示订货点为 L(即存储量为 L 时)而在 x 天内需求 $r>L$ 的概率。

$$F_x(L)=\sum_{r>L}\varphi_x(r)$$

P_L 的计算很繁，为了简化记缺货费的期望值为 C_2P_L，n_0 次缺货费的期望值为 $n_0C_2P_L$，每年的存储费用为$\left(\frac{1}{2}Q_0+B\right)C_1$。由于 Q_0 是根据存储费用和订购费用权衡后得出的最佳值，因此只需要考虑维持缓冲存储量的存储费用以及缺货费期望值两者之和最小。

即令 $n_0C_2P_L+C_1B$ 最小以确定 L 和 B。

由关系式 $L=\mu\rho+B$ 可知只要确定了 L 就相当于 B 也确定了。

例 12 (模型八)某厂生产中需用钢材，t 时间内需求的概率 $\varphi_t(r)$服从泊松分布：

$$\varphi_t(r)=\frac{e^{-\rho t}(\rho t)^r}{r!}$$

每天平均需求为 1 吨，$\rho=1$，年平均需求量 D 为 365 吨。则 t 时间内需求为 r 的概率

$$\varphi_t(r)=\frac{e^{-t}t^r}{r!}$$

备货时间为 x 天的概率服从正态分布

$$P(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}$$

平均备货时间 $\mu=15$(天)，方差 $\sigma^2=1$。

$$P(x)=\frac{1}{\sqrt{2\pi}}e^{-(x-15)^2/2}$$

年存储费用每吨为 50 元，每次订购费用为 1 500 元，缺货费用每吨为 5 000 元，问每年应分多少批次？又订购量 Q，缓冲存储量 B，订货点 L，各为何值才使费用最少？

解

$$Q_0=\sqrt{\frac{2\times365\times1\,500}{50}}\doteq146(\text{吨})$$

$$n_0=\frac{D}{Q_0}\doteq2.5(\text{次})$$

下面计算 L 及 B，各步算出的数值列于表 14-2。

表　14-2

(一) 备货时间 x	(二) 备货时间的概率 $P(x)$ $\left(P(x)=\frac{1}{\sqrt{2\pi}}e^{-(x-15)^2/2}\right)$	(三) x 天内的平均需求 (算式为 ρx)
① 13	0.05	13
② 14	0.24	14
③ 15	0.40	15
④ 16	0.24	16
⑤ 17	0.05	17

(四) 需求 $r>L$，$F_x(L)$ 的概率，算式 $F_x(L)=\sum\limits_{r>L}\frac{e^{-x}x^r}{r!}=1-\sum\limits_{r=0}^{L}\frac{e^{-x}x^r}{r!}$

$L=15$	$L=21$	$L=22$	$L=23$	$L=24$	$L=25$	$L=26$	$L=31$
① 0.236	0.014	0.008	0.004	0.002	0.001	0	0
② 0.331	0.029	0.017	0.009	0.005	0.003	0.001	0
③ 0.432	0.053	0.033	0.019	0.011	0.006	0.003	0
④ 0.533	0.089	0.058	0.037	0.022	0.013	0.008	0
⑤ 0.629	0.138	0.095	0.063	0.040	0.025	0.015	0.001

(五) 相应备货时间及需求两者概率的乘积 $P(x)F_x(L)$

(五)＝(二)×(四)

$L=15$	$L=21$	$L=22$	$L=23$	$L=24$	$L=25$	$L=26$	$L=31$
① 0.012	0	0	0	0	0	0	0
② 0.079	0.007	0.004	0.002	0.001	0.001	0	0
③ 0.173	0.021	0.013	0.008	0.004	0.002	0.001	0
④ 0.128	0.021	0.014	0.008	0.005	0.003	0.002	0
⑤ 0.031	0.007	0.005	0.003	0.002	0.001	0.001	0

根据表 14-2 算出 P_L、B 和费用的各种数值见表 14-3。

表 14-3

L	$L=15$	$L=21$	$L=22$	$L=23$	$L=24$	$L=25$	$L=26$	$L=31$
P_L	0.423	0.056	0.036	0.021	0.012	0.007	0.004	0
B	0	6	7	8	9	10	11	16
费用	5 287.5	1 000	800	663	600	588*	600	800

$$P_L = \sum_{x=15}^{17} P(x)F_x(L), B = L - \rho x$$

费用的计算式：

$$n_0C_2P_L + C_1B = 2.5 \times 5\,000 \times P_L + 50 \times B$$

说明 ① 备货时间小于 13，或大于 18 者，因为它们的概率很小，故略去。

② L 的选值可以多一些，如保证可以选到最小值，L 选值也可少一些。由表 14-3 中可以看到当 $L=25$，$B=10$ 费用 588* 为最小。据此即可确定存储策略。

答 该厂订购批量为 146 吨，订购点为 25 吨，每年订货 2.5 次(两年订货 5 次)，缓冲存储量为 10 吨。

当清点存储花费劳动多，或清点困难时，人们常把存储物分成三堆存放。以例 12 来说，将缓冲存储量 $B=10$ 吨放一处，称之为第三堆。将平均拖后时间内的平均需求量 $D_L=\mu\rho=15$ 吨放另一处称第二堆。第三堆、第二堆之和等于订货点 25 吨。其余存储另放一处称第一堆。平日从第一堆取用，第一堆用完，动用第二堆时，立即订货。动用第三堆时，即需采取措施以防缺货。

14.4 其他类型存储问题

有些存储问题远较本章所述模型复杂，上述公式不能用来求解，也可以利用运筹学的其他方法求解。如水库储水的调度问题，有人利用排队论方法处理问题，有人利用动态规划方法，都做出了成绩。

14.4.1 库容有限制的存储问题

下面介绍一个例题，与本章前述的方法无关，可用线性规划方法求解。

例 13 已知仓库最大容量为 A，原有存储量为 I，要计划在 m 个周期内，确定每一个周期的合理进货量与销售量，使总收入最多。已知第 i 个周期出售一个单位货物的收入为 a_i，而订购一个单位货物的订货费为 b_i，$i=1,2,\cdots,m$。

解 设 x_i，y_i 分别为第 i 个周期的进货量及售货量，这时总收入为

$$C = \sum_{i=1}^{m} (a_iy_i - b_ix_i)$$

要求出 x_i, y_i 使 C 达到最大值($i=1,2,\cdots,m$)。容易理解 x_i, y_i,这些变量不能任意取值。

(1) 它们受到库容的限制,即进货量加上原有存储量不能超过 A;

(2) 每个周期的售出量不能超过该周期的存储量;

(3) 进货量及售出量不能取负值。

用方程组表示上述的限制(约束条件):

(1) $I+\sum_{i=1}^{s}(x_i-y_i)\leqslant A \quad (s=1,2,\cdots,m)$

(2) $y_s\leqslant I+\sum_{i=1}^{s-1}(x_i-y_i) \quad (s=1,2,\cdots,m)$

(3) $x_i\geqslant 0, y_i\geqslant 0 \quad (i=1,2,\cdots,m)$

目标函数,约束条件都是线性的,可利用线性规划方法求解。

14.4.2 两级供应链订货量协调决策模型

1. 基于退货价格契约的两级供应链订货量协调决策模型

考虑一个由供应商和零售商组成的一个两级供应链(见图 14-12),假设某种原材料(零部件)的需求量 x 是连续的随机变量,密度函数为 $f(x)$,分布函数为 $F(x)$,供应商对这种原材料给出的批发价格为 W,原材料的单位成本为 C_S。零售商对这种原材料的销售价格为 P,单位成本(包括运输费用等)为 C_I。若零售商未售出的原材料以退货价格 S(也称单位原材料的残值)退还给供应商。零售商需要确定订货量 Q,使其期望利润最大。

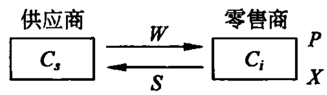


图 14-12 由供应商和零售商组成的一个两级供应链

(1) 零售商订货量决策

当 $x<Q$ 时,销售量为 x;当 $x\geqslant Q$ 时,销售量为 Q,故销售量函数为:

$$g(x)=\begin{cases}x, & x<Q\\ Q, & x\geqslant Q\end{cases} \tag{14-32}$$

因此,零售商的总期望利润为:

$$\begin{aligned}E(\pi_I)=&\int_0^Q(P-W-C_I)xf(x)\mathrm{d}x+\int_0^Q(S-W-C_I)(Q-x)f(x)\mathrm{d}x+\\&\int_Q^{+\infty}(P-W-C_I)Qf(x)\mathrm{d}x\\=&(P-W-C_I)Q-(P-S)\int_0^Q F(x)\mathrm{d}x\end{aligned} \tag{14-33}$$

$$\frac{\mathrm{d}E(\pi_I)}{\mathrm{d}Q}=(P-W-C_I)-(P-S)F(Q) \tag{14-34}$$

因为

$$\frac{\mathrm{d}^2 E(\pi_I)}{\mathrm{d}Q^2} = -(P-S)f(x) \leqslant 0 \tag{14-35}$$

所以 $E(\pi_I)$有最大值

令$\frac{\mathrm{d}E(\pi)}{\mathrm{d}Q}=0$,得

$$F(Q) = \frac{P-W-C_I}{P-S} \tag{14-36}$$

即零售商的最优订货量

$$Q^0 = F^{-1}\left[\frac{P-W-C_I}{P-S}\right] \tag{14-37}$$

将 Q^0 代入式(14-33)可得零售商的最优期望利润为

$$E(\pi_I^0) = (P-W-C_I)Q^0 - (P-S)\int_0^{Q_0} F(x)\mathrm{d}x \tag{14-38}$$

此时,供应商的最优期望利润可表示为

$$E(\pi_S^0) = (W-C_S)Q^0 \tag{14-39}$$

(2) 订货量协调决策模型

由式(14-37)可知,零售商根据供应商给出的批发价格和退货价格及原材料的单位成本和市场需求情况来确定最优订货量 Q^0,为了使该订货量 Q^0 与整个供应链的订货量一致,且使供应商与零售商双方均达到最优利润,须协商原材料的批发价格或退货价格。

a. 供应链的订货量

供应链的期望利润为

$$E(\pi_T) = P\left(Q - \int_0^Q F(x)\mathrm{d}x\right) - (C_I + C_S)Q \tag{14-40}$$

令$\frac{\mathrm{d}E(\pi_T)}{\mathrm{d}Q}=P(1-F(Q))-(C_I+C_S)=0$,得

$$F(Q) = \frac{P-C_I-C_S}{P} \tag{14-41}$$

因此,对供应链来说,最优订货量为:

$$Q^* = F^{-1}\left[\frac{P-C_I-C_S}{P}\right] \tag{14-42}$$

b. 退货价格

当 $Q^0=Q^*$ 时,零售商的最优订货量也是供应链整体运作绩效达到最优时的订货量,此时有:

$$\frac{P-W-C_I}{P-S} = \frac{P-C_I-C_S}{P} \tag{14-43}$$

即:

$$S = \frac{P(W-C_S)}{P-(C_I+C_S)} \tag{14-44}$$

c. 零售商的订货量

当原材料的批发价格 W 和退货价格 S 满足式(14-44)时，零售商的订货量与供应链的订货量一致。此时，可由式(14-37)或式(14-42)确定零售商的最优订货量。

例 14 设某原材料的市场价格为 $P=80$ 元，需求量 x 服从均值为 100，方差为 20^2 的正态分布，零售商的原材料单位处理成本 $C_I=10$，供应商的单位原材料获得成本为 $C_S=50$，批发价格 $W=60$，则由式(14-44)得原材料的退货价格为 $S=40$，由式(14-37)得零售商的最优订货量为 $Q^0=86$。

2. 基于收益分享契约的两阶段供应链协调模型

在一个由供应商和零售商组成的两阶段供应链中，收益分享(revenue sharing)契约由供应商的批发价格和供应商与零售商之间的收入分配比例两个参数来描述，供应商以低于其边际成本的批发价格将原材料(零部件)出售给零售商，同时分配得到一定比例的零售商的销售收入。收益分享契约的目的就是通过降低批发价格来激励零售商采用有利于供应链的总收益的订货量，使供应链的总期望利润与集中决策时的供应链的总期望利润相等，从而实现协调供应链的目的。

(1) 模型的描述

考虑一个由供应商和零售商构成的两级供应链。零售商 I 向供应商 S 订货的数量是 q，其需求量 x 是随机变量，分布密度函数为 $f(x)$。产品的销售价格是 p。C_S，C_I 分别为供应商 S 和零售商 I 的单位运营成本，供应商 S 提供给零售商 I 的批发价格是 ω，单位运输成本为 C_T，零售商 I 得到产品销售收入的比例是 ϕ，其余的 $(1-\phi)$ 分配给供应商 S。如图 14-13 所示。

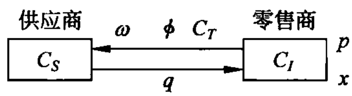


图 14-13 由供应商和零售商构成的两级供应链

用 Π_{IC}，Π_{SC} 分别表示零售商 I 和供应商 S 获得的利润，Π_T 表示整个供应链的总利润。则：

$$\Pi_{IC} = \phi p\min\{q,d\} - (\omega + C_I + C_T)q \tag{14-45}$$

$$\Pi_{SC} = (1-\phi)p\min\{q,d\} + (\omega - C_S)q \tag{14-46}$$

$$\Pi_T = p\min\{q,d\} - (C_S + C_I + C_T)q \tag{14-47}$$

零售商 I 和供应商 S 的期望利润分别为：

$$E(\Pi_{IC}) = \phi p\left[q - \int_0^q F(x)\mathrm{d}x\right] - (\omega + C_I + C_T)q \tag{14-48}$$

$$E(\Pi_{SC}) = (1-\phi)p\left[q - \int_0^q F(x)\mathrm{d}x\right] + \omega q - C_S q \tag{14-49}$$

$$E(\Pi_T) = p\left[q - \int_0^q F(x)\mathrm{d}x\right] - (C_I + C_S + C_T)q \tag{14-50}$$

分别对式(14-48)～式(14-50)两边求导，令其等于0，并解之，得零售商 I 的最优订货量 q_I、供应商的最优发货量 q_S 和供应链的最优产量 q_T 分别为

$$q_I = F^{-1}\left[\frac{\phi p - (\omega + C_I + C_T)}{\phi p}\right] \tag{14-51}$$

$$q_s = F^{-1}\left[\frac{(1-\phi)p - (C_S - \omega)}{(1-\phi)p}\right] \tag{14-52}$$

$$q_T = F^{-1}\left[\frac{p - (C_I + C_S + C_T)}{p}\right] \tag{14-53}$$

(2) 收益分享契约参数满足的条件

a. 渠道协调条件

收益分享契约实现供应链协调的手段就是为供应链中各节点企业提供一种协调约定，使各个阶段的决策者能够采取与供应链的最优策略相一致的策略。该契约应该使零售商的最优订货量 q_I、供应商的最优发货量 q_S 和供应链的最优产量 q_T 相等，即 $q_I = q_S = q_T$，使各阶段的最优策略和供应链的总的最优策略一致。

由 $q_I = q_T$ 或 $q_S = q_T$ 及式(14-51)～式(14-53)可以得到

$$\omega = \phi(C_I + C_T + C_S) - (C_I + C_T) \tag{14-54}$$

由于 ω 是产品的批发价格，所以必须满足 $\omega \geqslant 0$，由式(14-54)可得

$$\phi \geqslant \frac{C_I + C_T}{C_I + C_T + C_S} \tag{14-55}$$

b. 利润分配条件

供应商和零售商的期望利润必须大于其独立决策时相应的利润。因为只有这样，供应商和零售商才会自觉地履行收益分享契约的约定。

① 独立决策时的最优期望利润

假定供应商先确定批发价格 ω_0，零售商据此确定最优订货量 q_0，则此时零售商的最优期望利润为

$$E(\Pi_{I0}^*) = p\left[q_0 - \int_0^{q_0} F(x)\mathrm{d}x\right] - (\omega_0 + C_I + C_T)q_0 \tag{14-56}$$

式中

$$q_0 = F^{-1}\left[\frac{p - (\omega_0 + C_I + C_T)}{p}\right] \tag{14-57}$$

供应商的最优期望利润为

$$E(\Pi_{S0}^*) = (\omega_0 - C_S)q_0 \tag{14-58}$$

② 协调决策时的最优期望利润

将 q_T 代入式(14-50)得协调决策时供应链最优期望利润为

$$E(\Pi_T^*) = p\left[q_T - \int_0^{q_T} F(x)\mathrm{d}x\right] - (C_I + C_S + C_T)q_T \tag{14-59}$$

将 q_T 及式(14-54)代入式(14-48)和式(14-49)得协调决策时的零售商和供应商的最优期望利润分别为

$$E(\Pi_{IC}^*)=\phi p[q_T-\int_0^{q_T}F(x)\mathrm{d}x]-\phi(C_I+C_S+C_T)q_T \tag{14-60}$$

$$E(\Pi_{SC}^*)=(1-\phi)p[q_T-\int_0^{q_T}F(x)\mathrm{d}x]-(1-\phi)(C_I+C_S+C_T)q_T \tag{14-61}$$

将式(14-59)代入式(14-60)和式(14-61)得：

$$E(\Pi_{IC}^*)=\phi E(\Pi_T^*) \tag{14-62}$$

$$E(\Pi_{SC}^*)=(1-\phi)E(\Pi_T^*) \tag{14-63}$$

③ 契约参数满足的利润分配条件

由 $E(\Pi_{IC}^*)\geqslant E(\Pi_{I0}^*)$ 和 $E(\Pi_{SC}^*)\geqslant E(\Pi_{S0}^*)$ 得契约参数满足的利润分配条件为：

$$\phi\geqslant\frac{E(\Pi_{I0}^*)}{E(\Pi_T^*)} \tag{14-64}$$

$$\phi\leqslant 1-\frac{E(\Pi_{S0}^*)}{E(\Pi_T^*)} \tag{14-65}$$

由式(14-55)、式(14-64)和式(14-65)得 ϕ 的取值范围为

$$\max\left\{\frac{C_I+C_T}{C_I+C_T+C_S},\frac{E(\Pi_{I0}^*)}{E(\Pi_T^*)}\right\}\leqslant\phi\leqslant 1-\frac{E(\Pi_S^*)}{E(\Pi_T^*)} \tag{14-66}$$

(3) 收益分享契约参数取值过程

收益分享契约参数取值步骤如下：

① 由式(14-53)和式(14-57)确定 q_T 和 q_0；

② 由式(14-56)、式(14-58)和式(14-59)计算 $E(\Pi_{I0}^*)$、$E(\Pi_{S0}^*)$ 和 $E(\Pi_T^*)$；

③ 由式(14-66)计算 ϕ 的范围并取其值；

④ 由式(14-54)计算 ω；

⑤ 由式(14-62)和式(14-63)计算不同收益分享契约参数下的利润分配。

例 15 我们将用一个实例来阐明基于收益分享契约的利润分配。假设在上述模型中，x 服从 $N(110,20^2)$，其他参数的取值如表 14-4 所示。

表 14-4

变量	C_I	C_S	C_T	ω_0	p
值	10	28	8	45	80

根据上述契约参数确定过程中的步骤①和步骤②得协调决策和独立决策时的最优订货量分别为 $q_T=106$ 和 $q_0=94$，最优期望利润分别为 $E(\Pi_{I0}^*)=1\,406$、$E(\Pi_{S0}^*)=1\,599$ 和 $E(\Pi_T^*)=3\,113$。

由步骤③得 $0.452\leqslant\phi\leqslant0.486$，任取 $\phi=0.46$、0.47 和 0.48，则由步骤④求出相应的

ω,从而获得了收益分享契约的三组值(如表 14-5 的第 2 列至第 3 列)。由步骤⑤计算出的不同收益分享契约参数下的利润分配如表 14-5 第 4 列至第 5 列。

表 14-5 的第 3 行至第 5 行表示采用收益分享契约时的三组契约参数取值及相应的利润分配,第 6 行为独立决策时的利润分配。从表 14-5 可以看出,不同的契约参数取值其利润分配是不同的(但都比独立决策时的利润大)。

表 14-5

收益分享契约			不同契约下总期望利润的分配	
	ϕ	ω	供应商	零售商
(1)	0.46	3.16	1 681	1 432
(2)	0.47	3.62	1 650	1 463
(3)	0.48	4.08	1 619	1 494
独立决策时的利润			1 599	1 406

注 记

本章仅就存储问题作了简单的介绍。由于实际问题多种多样,存储模型也远比我们介绍过的模型多,解决存储问题的方法也比本章介绍过的方法多。例如多阶段存储问题就要利用动态规划方法求解(见本书"动态规划")。随机型存储问题,有时需要用电子计算机利用模拟技术求解。其他与存储有类似性质的问题可以用本章介绍的方法求解。例如制造特殊设备,其零部件只适用于这种设备,只能在制造设备时同时生产,否则制造费用要提高很多。这时就要考虑这些零部件要不要留备用件?备用件的数量是多少?这样的问题可以利用报童问题求解方法解决。

现在我国企业都很注意提高经济效益,因而存储问题引起企业的注意。"A、B、C 分类管理法"已被许多企业采用,进而研究利用电子计算机管理库存的问题。随着企业管理水平的提高,存储问题也更加受到人们的重视。为解决实践中提出的问题,进一步研究新的存储模型,这样便会促使存储论不断的发展。

习题

14.1 设某工厂每年需用某种原料 1 800 吨,不需每日供应,但不得缺货。设每吨每月的保管费为 60 元,每次订购费为 200 元,试求最佳订购量。

14.2 某公司采用无安全存量的存储策略。每年使用某种零件 100 000 件,每件每年的保管费用为 30 元,每次订购费为 600 元,试求:

(1) 经济订购批量。

(2) 订购次数。

14.3 设某工厂生产某种零件,每年需要量为 18 000 个,该厂每月可生产 3 000 个,每次生产的装配费为 5 000 元,每个零件的存储费为 1.5 元,求每次生产的最佳批量。

14.4 某产品每月用量为 4 件，装配费为 50 元，存储费每月每件为 8 元，求产品每次最佳生产量及最小费用。若生产速度为每月生产 10 件，求每次最佳生产量及最小费用。

14.5 每月需要某种机构零件 2 000 件，每件成本 150 元，每年的存储费用为成本的 16%，每次订购费 100 元，求 E. O. Q 及最小费用。

14.6 在 14.5 题中如允许缺货，求库存量 s 及最大缺货量，设缺货费为 $C_2=200$ 元。

14.7 某制造厂每周购进某种机构零件 50 件，订购费为 40 元，每周保管费为 3.6 元。

(1) 求 E. O. Q。

(2) 该厂为少占用流动资金，希望存储量达到最低限度，决定可使总费用超过最低费用的 4%作为存储策略，问这时订购批量为多少？

14.8 某公司采用无安全存量的存储策略，每年需电感 5 000 个，每次订购费 500 元，保管费用每年每个 10 元，不允许缺货。若采购少量电感每个单价 30 元，若一次采购 1 500 个以上则每个单价 18 元，问该公司每次应采购多少个？

（提示：本题属于订货量多，价格有折扣的类型，即订货费 C_3+KQ，K 为阶梯函数。）

14.9 某工厂的采购情况为

采购数量/单位	单价/元
0～1 999	100
2 000 以上	80

假设年需要量为 10 000 单位，每次订货费为 2 000 元，存储费率为 20%，则每次应采购多少？

14.10 一个允许缺货的 E. O. Q 模型的费用绝不会超过一个具有相同存储费、订购费，但不允许缺货的 E. O. Q 模型的费用，试说明之。

14.11 某厂对原料需求的概率如表 14-6：

表 14-6

需求量 r/吨	20	30	40	50	60
概率 $P(r)(\sum P(r)=1)$	0.1	0.2	0.3	0.3	0.1

每次订购费 $C_3=500$ 元，原料每吨价 $K=400$ 元，每吨原料存储费 $C_1=50$ 元，缺货费每吨 $C_2=600$ 元。该厂希望制定(s,S)型存储策略，试求 s 及 S 值。

14.12 某厂需用配件数量 r 是一个随机变量，其概率服从泊松分布。t 时间内需求概率为

$$\varphi_t(r)=\frac{e^{-\rho t}(\rho t)^r}{r!} \quad \text{平均每日需求为 } 1(\rho=1)$$

备货时间为 x 天的概率服从正态分布

$$P(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}$$

平均拖后时间 $\mu=14$ 天，方差 $\sigma^2=1$，在生产循环周期内存储费 $C_1=1.25$ 元，缺货费 $C_2=10$ 元，装配费 $C_3=3$ 元。问两年内应分多少批订货？每次批量及缓冲存储量各为何值才能使总费用最小？

参考资料

[1] 林文源. 物料管理学. 澳门：澳门科技丛书出版社，1978
[2] 吴云从. 随机存储的几个问题. 系统工程(第一期)，1984
[3] 严颖，程世学，程侃. 运筹学随机模型. 北京：中国人民大学出版社，1995

Part 9 第9篇

对策论

对策也叫博弈，是自古以来的政治家和军事家都很注意研究的问题。作为一门正式学科，是在20世纪40年代形成并发展起来的。直到1944年冯·诺依曼(von Neumann)与摩根斯特恩(O. Morgenstern)的《博弈论与经济行为》一书出版，标志着现代博弈论的初步形成。书中提出的标准型、扩展型和合作型博弈模型解的概念和分析方法，奠定了这门学科的理论基础，成为使用严谨的数学模型研究冲突对抗条件下最优决策问题的理论。然而，诺依曼的博弈论的局限性也日益暴露出来。由于它过于抽象，使应用范围受到很大限制，所以影响力很有限。20世纪50年代，纳什(Nash)建立了非合作博弈的“纳什均衡”理论，标志着博弈论的新时代开始。正是因为纳什在经济博弈论领域划时代的贡献，使其成为继冯·诺依曼之后对对策论发展贡献最大的学者，并因此获得了1994年诺贝尔经济学奖。纳什均衡概念在非合作博弈理论中起着核心作用，为博弈论广泛应用于经济学、管理学、社会学、政治学、军事科学等领域奠定了坚实的理论基础。

CHAPTER 15
第15章

对策论基础

15.1 引言

15.1.1 对策行为和对策论

对策论亦称竞赛论或博弈论，是研究具有斗争或竞争性质现象的数学理论和方法。一般认为，它是现代数学的一个新分支，是运筹学的一个重要学科。对策论发展的历史并不长，但由于它研究的问题与政治、经济、军事活动乃至一般的日常生活等有着密切联系，并且其分析和解决问题的方法具有明显特色，所以日益引起广泛重视。

在日常生活中，经常会看到一些相互之间具有斗争或竞争性质的行为，如下棋、打牌、体育比赛等。又比如战争活动中的双方，都力图选取对自己最有利的策略，千方百计去战胜对手。还比如国际间的谈判，各种政治力量之间的斗争，各国际集团之间的斗争等，无一不具有斗争的性质。在经济领域，各国之间的谈判、企业之间的谈判、企业之间为争夺市场而进行的竞争等，举不胜举。

具有斗争或竞争性质的行为称为对策行为。在这类行为中，参加斗争或竞争的各方各自具有不同的目标和利益。为了达到各自的目标和利益，各方必须考虑对手的各种可能的行动方案，并力图选取对自己最有利或最合理的方案。对策论就是研究在对策行为中，斗争各方是否存在最有利或最合理的行动方案，以及如何找到最有利或最合理行动方案的数学理论和方法。

在我国古代，“齐王赛马”就是一个典型的对策论的例子。

战国时期，有一天齐王提出要与田忌赛马，双方约定从各自的上、中、下三个等级的马中各选一匹参赛，每匹马均只能参赛一次，每一次比赛双方各出一匹马，负者要付给胜者千金。已经知道的是，在同等级的马中，田忌的马不如齐王的马，但如果田忌的马比齐王的马高一等级，则田忌的马可取胜。当时，田忌手下的一个谋士给他出了个主意：每次比赛时先让齐王牵出他要参赛的马，然后用下马对齐王的上马，用中马对齐王的下马，用上马对齐王的中马。比赛结果，田忌二胜一负，夺得千金。由此看来，两个人各采取什么样的出马次序对胜负是至关重要的。

15.1.2 对策行为的三个基本要素

1. 局中人

在一个对策行为(或一局对策)中,有权决定自己行动方案的对策参加者称为局中人。通常用 I 表示局中人的集合。如果有 n 个局中人,则 $I=\{1,2,\cdots,n\}$。一般要求一个对策中至少要有两个局中人。如在"齐王赛马"的例子中,局中人是齐王和田忌。

2. 策略集

一局对策中,可供局中人选择的一个实际可行的完整的行动方案称为一个策略。参加对策的每一局中人 $i,i\in I$,都有自己的策略集 S_i。一般,每一局中人的策略集中至少应包括两个策略。

在"齐王赛马"的例子中,如果用(上,中,下)表示以上马、中马、下马依次参赛这样一个次序,这就是一个完整的行动方案,即为一个策略。可见,局中人齐王和田忌各自都有6个策略:(上、中、下),(上、下、中),(中、上、下),(中、下、上),(下、中、上),(下、上、中)。

3. 赢得函数(支付函数)

在一局对策中,各局中人选定的策略形成的策略组称为一个局势,即若 S_i 是第 i 个局中人的一个策略,则 n 个局中人的策略组

$$s=(s_1,s_2,\cdots,s_n)$$

就是一个局势。全体局势的集合 S 可用各局中人策略集的笛卡儿积表示,即

$$S=S_1\times S_2\times\cdots\times S_n$$

当一个局势出现后,对策的结果也就确定了。也就是说,对任一局势 $s\in S$,局中人 i 可以得到一个赢得值 $H_i(s)$。显然,$H_i(s)$ 是局势 s 的函数,称为第 i 个局中人的赢得函数。在齐王与田忌赛马的例子中,局中人集合为 $I=\{1,2\}$,齐王和田忌的策略集可分别用 $S_1=\{\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5,\alpha_6\}$ 和 $S_2=\{\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,\beta_6\}$ 表示。这样,齐王的任一策略 α_i 和田忌的任一策略 β_j 就形成了一个局势 s_{ij}。如果 $\alpha_1=$(上,中,下),$\beta_1=$(上,中,下),则在局势 s_{11} 下齐王的赢得值为 $H_1(s_{11})=3$,田忌的赢得值为 $H_2(s_{11})=-3$,如此等等。

以上讨论了局中人、策略集和赢得函数这三个概念。当这三个基本要素确定后,一个对策模型也就给定了。

15.1.3 对策问题举例及对策的分类

对策论在经济管理的众多领域中有着十分广泛的应用,下面列举几个可以用对策论思想和模型进行分析的例子。

例 1 (市场购买力争夺问题)据预测,某乡镇下一年的饮食品购买力将有4 000万元。乡镇企业和中心城市企业饮食品的生产情况是:乡镇企业有特色饮食品和低档饮食品两类,中心城市企业有高档饮食品和低档饮食品两类产品。它们争夺这一部分购买力

的结局见表15-1(表中数字的单位是万元),问题是乡镇企业和中心城市企业应如何选择对自己最有利的产品策略。

表15-1 乡镇企业所得

乡镇企业的策略	中心城市企业的策略	
	出售高档饮食品	出售低档饮食品
出售特色饮食品	2 000	3 000
出售低档饮食品	1 000	3 000

例2 (销售竞争问题)假定企业Ⅰ,Ⅱ均能向市场出售某一产品,不妨假定它们可于时间区间[0,1]内任一时点出售。设企业Ⅰ在时刻x出售,企业Ⅱ在时刻y出售,则企业Ⅰ的收益(赢得)函数为

$$H(x,y)=\begin{cases} c(y-x) & 若\ x<y \\ \frac{1}{2}c(1-x) & 若\ x=y \\ c(1-x) & 若\ x>y \end{cases}$$

问这两个企业各选择什么时机出售对自己最有利?在这个例子中,企业Ⅰ,Ⅱ可选择的策略均有无穷多个。

例3 (费用分摊问题)假设沿某一河流有相邻的3个城市A、B、C,各城市可单独建立水厂,也可合作兴建一个大水厂。经估算,合建一个大水厂,加上敷设管道的费用,要比单独建3个小水厂的总费用少。但合建大厂的方案能否实施,要看总的建设费用能否在3个城市之间进行合理的分摊。如果某个城市分摊到的费用比它单独建设水厂的费用还多的话,它显然不会接受合作的方案。因此,需要研究的问题是:如何合理地分摊总的建设费用,使合作兴建大水厂的方案得以实现?

例4 (拍卖问题)最常见的一种拍卖形式是先由拍卖商把拍卖品描述一番,然后提出第一个报价。接下来由买者报价,每一次报价都要比前一次高,最后谁出的价最高拍卖品即归谁所有。假设有n个买主给出的报价分别为$p_1,\cdots,p_n$,且不妨设$p_n>p_{n-1}>\cdots>p_1$,则买主n只要报价略高于p_{n-1},就能买到拍卖品,即拍卖品实际上是在次高价格上卖出的。现在的问题是,各买主之间可能知道他人的估价,也可能不知道他人的估价,每人应如何报价对自己能以较低的价格得到拍卖品最为有利?最后的结果又会怎样?

例5 (囚犯难题)设有两个嫌疑犯因涉嫌作案被警官拘留,警官分别对两人进行审讯。根据法律,如果两个人都承认此案是他们干的,则每人各判刑7年;如果两人都不承认,则由于证据不足,两人将各获刑1年;如果只有一人承认并揭发对方,则承认者予以宽大释放,而不承认者将判刑9年。因此,对两个囚犯来说,面临着一个在“承认”和“不承认”这两个策略间进行选择的难题。

上面几个例子都可看成是一个对策问题,所不同的是有些是二人对策,有些是多人对策;有些是有限对策,有些是无限对策;有些是零和对策,有些是非零和对策;有些是合

作对策，有些是非合作对策，等等。为了便于对不同的对策问题进行研究，可以根据不同方式进行分类，通常的分类方式有：

(1) 根据局中人的个数，分为二人对策和多人对策；

(2) 根据各局中人的赢得函数的代数和是否为零，分为零和对策与非零和对策；

(3) 根据各局中人间是否允许合作，分为合作对策和非合作对策；

(4) 根据局中人策略集中的策略个数，分为有限对策和无限对策。

此外，还有许多其他的分类方式。例如根据策略的选择是否与时间有关，可分为静态对策和动态对策；根据对策模型的数学特征，可分为矩阵对策、连续对策、微分对策、阵地对策、凸对策、随机对策等。

在众多对策模型中，占有重要地位的是二人有限零和对策(finite two-person zerosum game)，又称为矩阵对策。这类对策是到目前为止在理论研究和求解方法方面都比较完善的一个对策。尽管矩阵对策是一类最简单的对策模型，但其研究思想和方法十分具有代表性，足以体现对策论的一般思想和分析方法，且矩阵对策的基本结果也是研究其他对策模型的基础。基于上述原因，本章将着重介绍矩阵对策的基本内容，而只对其他对策模型作简要介绍。

15.2 矩阵对策的基本定理

15.2.1 矩阵对策的数学模型

矩阵对策即为二人有限零和对策，是指只有两个参加对策的局中人，每个局中人都只有有限个策略可供选择。在任一局势下，两个局中人的赢得之和总是等于零，即双方的利益是激烈对抗的，一方的所得即为另一方的所失。“齐王赛马”就是一个矩阵对策的例子，齐王和田忌各有 6 个策略，一局对策结束后，齐王的所得必为田忌的所失，反之亦然。

在矩阵对策中，一般用Ⅰ、Ⅱ分别表示两个局中人，并设局中人Ⅰ有 m 个纯策略(以与后面的混合策略区别)$\alpha_1,\alpha_2,\cdots,\alpha_m$，局中人Ⅱ有 n 个纯策略 $\beta_1,\beta_2,\cdots,\beta_n$，则局中人Ⅰ、Ⅱ的策略集分别为

$$S_1=\{\alpha_1,\alpha_2,\cdots,\alpha_m\}$$

$$S_2=\{\beta_1,\beta_2,\cdots,\beta_n\}$$

当局中人Ⅰ选定纯策略 α_i 和局中人Ⅱ选定纯策略 β_j 后，就形成了一个纯局势 (α_i,β_j)。可见这样的纯局势共有 $m\times n$ 个。对任一纯局势 (α_i,β_j)，记局中人Ⅰ的赢得值为 a_{ij}，并称

$$\boldsymbol{A}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

为局中人Ⅰ的赢得矩阵(或为局中人Ⅱ的支付矩阵)。由于假定对策为零和的,故局中人Ⅱ的赢得矩阵就是$-\boldsymbol{A}^{\mathrm{T}}$。

当局中人Ⅰ、Ⅱ和策略集S_1、S_2及局中人Ⅰ的赢得矩阵$\boldsymbol{A}$确定后,一个矩阵对策也就给定了。通常,将一个矩阵对策记成

$$G=\{\mathrm{I},\mathrm{II};S_1,S_2;\boldsymbol{A}\}\quad 或\quad G=\{S_1,S_2;\boldsymbol{A}\}$$

在"齐王赛马"的例子中齐王的赢得可用表15-2表示。[①]

表 15-2 齐王的赢得

齐王的策略	田忌的策略					
	$[\beta_1]$(上、中、下)	$[\beta_2]$(上、下、中)	$[\beta_3]$(中、上、下)	$[\beta_4]$(中、下、上)	$[\beta_5]$(下、中、上)	$[\beta_6]$(下、上、中)
α_1(上、中、下)	3	1	1	1	1	−1
α_2(上、下、中)	1	3	1	1	−1	1
α_3(中、上、下)	1	−1	3	1	1	1
α_4(中、下、上)	−1	1	1	3	1	1
α_5(下、中、上)	1	1	−1	1	3	1
α_6(下、上、中)	1	1	1	−1	1	3

赢得矩阵为

$$\boldsymbol{A}=\begin{bmatrix}3&1&1&1&1&-1\\1&3&1&1&-1&1\\1&-1&3&1&1&1\\-1&1&1&3&1&1\\1&1&-1&1&3&1\\1&1&1&-1&1&3\end{bmatrix}$$

当矩阵对策模型给定后,各局中人面临的问题便是如何选取对自己最有利的纯策略,以谋取最大的赢得(或最少损失)。下面通过一个具体例子来分析应如何求解各局中人的最优纯策略。

例 6 设有一矩阵对策$G=\{S_1,S_2;\boldsymbol{A}\}$,其中$S_1=\{\alpha_1,\alpha_2,\alpha_3,\alpha_4\}$,$S_2=\{\beta_1,\beta_2,\beta_3\}$,

$$\boldsymbol{A}=\begin{bmatrix}-6&1&-8\\3&2&4\\9&-1&-10\\-3&0&6\end{bmatrix}$$

① 中国学者张盛开教授首先用矩阵对策模型来研究"齐王赛马"的问题,给出了齐王的赢得矩阵。(见《矩阵对策初步》,上海教育出版社,1980)

对于一般矩阵对策,有如下定义。

定义 1 设 $G=\{S_1,S_2;\boldsymbol{A}\}$为矩阵对策。其中

$$S_1=\{\alpha_1,\alpha_2,\cdots,\alpha_m\},S_2=\{\beta_1,\beta_2,\cdots,\beta_n\},\boldsymbol{A}=(a_{ij})_{m\times n}$$

若等式

$$\max_i \min_j a_{ij}=\min_j \max_i a_{ij}=a_{i^*j^*} \tag{15-1}$$

成立,记 $V_G=a_{i^*j^*}$。则称 V_G 为对策 G 的值,称使式(15-1)成立的纯局势$(\alpha_{i^*},\beta_{j^*})$为 G 在纯策略下的解(或平衡局势),α_{i^*} 与 β_{j^*} 分别称为局中人Ⅰ,Ⅱ的最优纯策略。

由定义 1 可知,在矩阵对策中两个局中人都采取最优纯策略(如果最优纯策略存在)才是理智的行动。

例 7 求解矩阵对策 $G=\{S_1,S_2;\boldsymbol{A}\}$,其中

$$\boldsymbol{A}=\begin{bmatrix}-7 & 1 & -8\\ 3 & 2 & 4\\ 16 & -1 & -3\\ -3 & 0 & 5\end{bmatrix}$$

解 根据矩阵 $\boldsymbol{A}$,有表 15-3。

表 15-3

	β_1	β_2	β_3	$\min_j a_{ij}$
α_1	−7	1	−8	−8
α_2	3	2	4	2*
α_3	16	−1	−3	−3
α_4	−3	0	5	−3
$\max_i a_{ij}$	16	2*	5	

于是

$$\max_i \min_j a_{ij}=\min_j \max_i a_{ij}=a_{22}=2$$

由定义 1,$V_G=2$,G 的解为(α_2,β_2),α_2 与 β_2 分别是局中人Ⅰ和Ⅱ的最优纯策略。

从例 7 可以看出,矩阵 $\boldsymbol{A}$ 的元素 a_{22} 既是其所在行的最小元素,又是其所在列的最大元素,即

$$a_{i2}\leqslant a_{22}\leqslant a_{2j}\quad i=1,2,3,4;\quad j=1,2,3$$

将这一事实推广到一般矩阵对策,可得如下定理。

定理 1 矩阵对策 $G=\{S_1,S_2;\boldsymbol{A}\}$在纯策略意义下有解的充分必要条件是:存在纯局势$(\alpha_{i^*},\beta_{j^*})$使得对一切 $i=1,\cdots,m,j=1,\cdots,n$,均有

$$a_{ij^*}\leqslant a_{i^*j^*}\leqslant a_{i^*j} \tag{15-2}$$

证明 先证充分性，由于对任意 i,j 均有

$$a_{ij^*}\leqslant a_{i^*j^*}\leqslant a_{i^*j}$$

故

$$\max_i a_{ij^*}\leqslant a_{i^*j^*}\leqslant\min_j a_{i^*j}$$

又因

$$\min_j\max_i a_{ij}\leqslant\max_i a_{ij^*}$$

$$\min_j a_{i^*j}\leqslant\max_i\min_j a_{ij}$$

所以

$$\min_j\max_i a_{ij}\leqslant a_{i^*j^*}\leqslant\max_i\min_j a_{ij}\tag{15-3}$$

另一方面，对任意 i,j 有

$$\min_j a_{ij}\leqslant a_{ij}\leqslant\max_i a_{ij}$$

所以

$$\max_i\min_j a_{ij}\leqslant\min_j\max_i a_{ij}\tag{15-4}$$

由式(15-3)和式(15-4)有

$$\max_i\min_j a_{ij}=\min_j\max_i a_{ij}=a_{i^*j^*}$$

且

$$V_G=a_{i^*j^*}.$$

现在来证明必要性。设有 i^*,j^* 使得

$$\min_j a_{i^*j}=\max_i\min_j a_{ij}$$

$$\max_i a_{ij^*}=\min_j\max_i a_{ij}$$

则由

$$\max_i\min_j a_{ij}=\min_j\max_i a_{ij}$$

有

$$\max_i a_{ij^*}=\min_j a_{i^*j}\leqslant a_{i^*j^*}\leqslant\max_i a_{ij^*}=\min_j a_{i^*j}$$

所以对任意 i,j 有

$$a_{ij^*}\leqslant\max_i a_{ij^*}\leqslant a_{i^*j^*}\leqslant\min_j a_{i^*j}\leqslant a_{i^*j}$$

为了便于对更广泛的对策情形进行分析，现引进关于二元函数鞍点的概念。

定义 2 设 $f(x,y)$ 为一个定义在 $x\in A$ 及 $y\in B$ 上的实值函数，如果存在 $x^*\in A$，$y^*\in B$，使得对一切 $x\in A$ 和 $y\in B$，有

$$f(x,y^*)\leqslant f(x^*,y^*)\leqslant f(x^*,y)\tag{15-5}$$

则称 (x^*,y^*) 为函数 f 的一个鞍点。

由定义 2 及定理 1 可知，矩阵对策 G 在纯策略意义下有解，且 $V_G=a_{i^*j^*}$ 的充要条件是：$a_{i^*j^*}$ 是矩阵 $\boldsymbol{A}$ 的一个鞍点。在对策论中，矩阵 $\boldsymbol{A}$ 的鞍点也称为对策的鞍点。

关于定理1中式(15-2)的直观解释是：如果 $a_{i^*j^*}$ 既是矩阵 $\boldsymbol{A}=(a_{ij})_{m\times n}$ 中第 i^* 行的最小值，又是 $\boldsymbol{A}$ 中第 j^* 列的最大值，则 $a_{i^*j^*}$ 即为对策的值，且 $(\alpha_{i^*},\beta_{i^*})$ 就是对策的解。其对策意义是：一个平衡局势 $(\alpha_{i^*},\beta_{i^*})$ 应具有这样的性质，当局中人Ⅰ选取了纯策略 α_{i^*} 后，局中人Ⅱ为了使其所失最少，只有选择纯策略 β_{j^*}，否则就可能失得更多；反之，当局中人Ⅱ选取了纯策略 β_{j^*} 后，局中人Ⅰ为了得到最大的赢得也只能选取纯策略 α_{i^*}，否则就会赢得更少。双方的竞争在局势 $(\alpha_{i^*},\beta_{j^*})$ 下达到了一个平衡状态。

例 8 求对策的解。设矩阵对策 $G=\{S_1,S_2;\boldsymbol{A}\}$，其中 $S_1=\{\alpha_1,\alpha_2,\alpha_3,\alpha_4\}$，$S_2=\{\beta_1,\beta_2,\beta_3,\beta_4\}$，赢得矩阵为

$$\boldsymbol{A}=\begin{bmatrix}6&5&6&5\\1&4&2&-1\\8&5&7&5\\0&2&6&2\end{bmatrix}$$

解 直接在 $\boldsymbol{A}$ 提供的赢得表上计算，有

$$\begin{array}{c|cccc|c} & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \min\\ \alpha_1 & 6 & 5 & 6 & 5 & 5^*\\ \alpha_2 & 1 & 4 & 2 & -1 & -1\\ \alpha_3 & 8 & 5 & 7 & 5 & 5^*\\ \alpha_4 & 0 & 2 & 6 & 2 & 0\\ \max & 8 & 5^* & 7 & 5^* & \end{array}$$

于是

$$\max_i \min_j a_{ij}=\min_j \max_i a_{ij}=a_{i^*j^*}=5$$

其中

$$i^*=1,3;\quad j^*=2,4$$

故 (α_1,β_2)，(α_1,β_4)，(α_3,β_2)，(α_3,β_4) 四个局势都是对策的解，且 $V_G=5$。

由例8可知，一般矩阵对策的解可以是不唯一的。当解不唯一时，解之间的关系具有下面两条性质。

性质 1 无差别性。即若 $(\alpha_{i_1},\beta_{j_1})$ 和 $(\alpha_{i_2},\beta_{j_2})$ 是对策 G 的两个解，则 $a_{i_1j_1}=a_{i_2j_2}$。

性质 2 可交换性。即若 $(\alpha_{i_1},\beta_{j_1})$ 和 $(\alpha_{i_2},\beta_{j_2})$ 是对策 G 的两个解，则 $(\alpha_{i_1},\beta_{j_2})$ 和 $(\alpha_{i_2},\beta_{j_1})$ 也是解。

上面两条性质的证明留给读者作为练习。这两条性质表明，矩阵对策的解不必是唯一的，但对策的值是唯一的。即当局中人Ⅰ采用构成解的最优纯策略时，能保证他的赢得 V_G 不依赖于对方的纯策略。

下面，举一个实际应用的例子。

例 9 某单位采购员在秋天要决定冬季取暖用煤的储量问题。已知在正常的冬季气温条件下要消耗15吨煤，在较暖与较冷的气温条件下要消耗10吨和20吨。假定冬季时

的煤价随天气寒冷程度而有所变化，在较暖、正常、较冷的气候条件下每吨煤价分别为100元，150元和200元，又设秋季时煤价为每吨100元。在没有关于当年冬季准确的气象预报的条件下，问秋季储煤多少吨能使单位购煤的支出最少？

这个问题可以看成是一个对策问题，把采购员当做局中人Ⅰ，他有三个策略：在秋天时买10吨、15吨与20吨，分别记为$\alpha_1,\alpha_2,\alpha_3$。

把大自然看做局中人Ⅱ，大自然的“选择”是冬季的气温，也有三种策略：出现较暖的、正常的与较冷的冬季，分别记为β_1,β_2,β_3。

如果将该单位冬季取暖用煤的实际支出（即秋季购煤时的用费与冬季不够时再补购的费用之和）作为局中人Ⅰ的赢得，得采购员的赢得矩阵如下

$$
\begin{array}{c} \\ \alpha_1(10\text{ 吨}) \\ \alpha_2(15\text{ 吨}) \\ \alpha_3(20\text{ 吨}) \end{array}
\begin{array}{c} \beta_1(\text{较暖})\ \beta_2(\text{正常})\ \beta_3(\text{较冷}) \\ \begin{bmatrix} -1\,000 & -1\,750 & -3\,000 \\ -1\,500 & -1\,500 & -2\,500 \\ -2\,000 & -2\,000 & -2\,000 \end{bmatrix} \end{array}
$$

$$\max_i \min_j a_{ij} = \min_j \max_i a_{ij} = a_{33} = -200$$

故对策的解为(α_3,β_3)，即秋季储煤20吨最为合理。

15.2.2 矩阵对策的混合策略

由上节的讨论可知，对矩阵对策$G=\{S_1,S_2;\boldsymbol{A}\}$来说，局中人Ⅰ有把握的至少赢得是

$$v_1=\max_i \min_j a_{ij}$$

局中人Ⅱ有把握的至多损失是

$$v_2=\min_j \max_i a_{ij}$$

显然，局中人Ⅰ的赢得值不会多于局中人Ⅱ的所失值，即总有$v_1\leqslant v_2$。当$v_1=v_2$时，矩阵对策G存在纯策略意义下的解，且$V_G=v_1=v_2$。然而，一般情形不总是如此，实际中出现的更多情形是$v_1<v_2$，这样，根据定义1，对策G就不存在纯策略意义下的解。例如对于赢得矩阵为

$$\boldsymbol{A}=\begin{bmatrix} 3 & 6 \\ 5 & 4 \end{bmatrix}$$

的对策问题来说，可知

$$v_1=\max_i \min_j a_{ij}=4, i*=2$$

$$v_2=\min_j \max_i a_{ij}=5, j*=1$$

$$v_2=a_{21}=5>4=v_1$$

于是，当双方各根据从最不利情形中选取最有利结果的原则选择纯策略时，应分别选取α_2和β_1，此时局中人Ⅰ将赢得5，比其预期赢得$v_1=4$还多，原因就在于局中人Ⅱ选择了β_1，使他的对手多得了原来不该得的赢得。故β_1对局中人Ⅱ来说并不是最优的，因而

他会考虑出 β_2。局中人Ⅰ亦会采取相应的办法，改出 α_1 以使赢得为 6，而局中人Ⅱ又可能仍取策略 β_1 来对付局中人Ⅰ的策略 α_1。这样，局中人Ⅰ出 α_1 或 α_2 的可能性以及局中人Ⅱ出 β_1 或 β_2 的可能性都不能排除。对两个局中人来说，不存在一个双方均可接受的平衡局势，或者说当 $v_1 < v_2$ 时，矩阵对策 G 不存在纯策略意义下的解。

在这种情况下，一个比较自然且合乎实际的想法是：既然各局中人没有最优纯策略可出，是否可以给出一个选取不同策略的概率分布。如在上例中，局中人Ⅰ可以制定如下一种策略：分别以概率 1/4 和 3/4 选取纯策略 α_1 和 α_2，这种策略是局中人Ⅰ的策略集 $\{\alpha_1, \alpha_2\}$ 上的一个概率分布，称之为混合策略。同样，局中人Ⅱ也可制定这样一种混合策略：分别以概率 1/2、1/2 选取纯策略 β_1、β_2。下面给出矩阵对策混合策略的定义。

定义 3 设有矩阵对策 $G=\{S_1, S_2; \boldsymbol{A}\}$，其中 $S_1=\{\alpha_1, \alpha_2, \cdots, \alpha_m\}$，$S_2=\{\beta_1, \beta_2, \cdots, \beta_n\}$，$\boldsymbol{A}=(a_{ij})_{m\times n}$ 记

$$S_1^* = \left\{x \in E^m \mid x_i \geqslant 0, \quad i = 1, \cdots, m, \quad \sum_{i=1}^{m} x_i = 1\right\}$$

$$S_2^* = \left\{y \in E^n \mid y_j \geqslant 0, \quad j = 1, \cdots, n, \quad \sum_{j=1}^{n} y_j = 1\right\}$$

则 S_1^* 和 S_2^* 分别称为局中人Ⅰ和Ⅱ的混合策略集(或策略集)；$x \in S_1^*$ 和 $y \in S_2^*$ 分别称为局中人Ⅰ和Ⅱ的混合策略(或策略)；对 $x \in S_1^*$，$y \in S_2^*$，称 (x, y) 为一个混合局势(或局势)，局中人Ⅰ的赢得函数记成

$$E(x, y) = x^{\mathrm{T}} \boldsymbol{A} y = \sum_i \sum_j a_{ij} x_i y_j \tag{15-6}$$

这样得到的一个新的对策记成 $G^* = \{S_1^*, S_2^*, E\}$，称 G^* 为对策 G 的混合扩充。

由定义 3 可知，纯策略是混合策略的特例。例如局中人Ⅰ的纯策略 α_k 等价于混合策略 $x=(x_1, \cdots, x_m)^{\mathrm{T}} \in S_1^*$，其中

$$x_i = \begin{cases} 1 & i = k \\ 0 & i \neq k \end{cases}$$

一个混合策略 $x=(x_1, \cdots, x_m)^{\mathrm{T}}$ 可设想成当两个局中人多次重复进行对策 G 时，局中人Ⅰ分别采取纯策略 $\alpha_1, \cdots, \alpha_m$ 的频率。若只进行一次对策，混合策略 $x=(x_1, \cdots, x_m)^{\mathrm{T}}$ 可设想成局中人Ⅰ对各纯策略的偏爱程度。

下面讨论矩阵对策 G 在混合策略意义下解的定义。

设两个局中人仍像前面一样地进行有理智的对策。当局中人Ⅰ采取混合策略 x 时，他只能希望获得(最不利的情形)

$$\min_{y \in S_2^*} E(x, y)$$

因此局中人Ⅰ应选取 $x \in S_1^*$，使得上式取极大值(最不利当中的最有利情形)，即局中人Ⅰ可保证自己的赢得期望值不少于

$$v_1 = \max_{x \in S_1^*} \min_{y \in S_2^*} E(x, y) \tag{15-7}$$

同理，局中人Ⅱ可保证自己的所失期望值至多是

$$v_2=\min_{y\in S_2^*}\max_{x\in S_1^*}E(x,y) \tag{15-8}$$

首先，注意到式(15-7)和式(15-8)是有意义的。因为根据定义，局中人Ⅰ的赢得函数 $E(x,y)$ 是欧氏空间 E^{m+n} 内有界闭集 D 上的连续函数，其中

$$D=\{(x,y)\mid x_i\geqslant 0,y_j\geqslant 0,i=1,\cdots,m,j=1,\cdots,n,$$
$$\sum_i x_i=1,\sum_j y_j=1\}$$

因此，对固定的 x，$E(x,y)$ 是 S_2^* 上的连续函数，故 $\min\limits_{y\in S_2^*}E(x,y)$ 存在，而且 $\min\limits_{y\in S_2^*}E(x,y)$ 也是 S_1^* 上的连续函数，故 $\max\limits_{x\in S_1^*}\min\limits_{y\in S_2^*}E(x,y)$ 存在。同样可说明 $\min\limits_{y\in S_2^*}\max\limits_{x\in S_1^*}E(x,y)$ 存在。

其次，仍然有 $v_1\leqslant v_2$。事实上，设

$$\max_{x\in S_1^*}\min_{y\in S_2^*}E(x,y)=\min_{y\in S_2^*}E(x^*,y)$$
$$\min_{y\in S_2^*}\max_{x\in S_1^*}E(x,y)=\max_{x\in S_1^*}E(x,y^*)$$

于是

$$v_1=\min_{y\in S_2^*}E(x^*,y)\leqslant E(x^*,y^*)\leqslant\max_{x\in S_1^*}E(x,y^*)=v_2$$

定义 4 设 $G^*=\{S_1^*,S_2^*;E\}$ 是矩阵对策 $G=\{S_1,S_2;\boldsymbol{A}\}$ 的混合扩充，如果

$$\max_{x\in S_1^*}\min_{y\in S_2^*}E(x,y)=\min_{y\in S_2^*}\max_{x\in S_1^*}E(x,y) \tag{15-9}$$

记其值为 V_G。则称 V_G 为对策 G^* 的值，称使式(15-9)成立的混合局势 (x^*,y^*) 为 G 在混合策略意义下的解(或简称解)，x^* 和 y^* 分别称为局中人Ⅰ和Ⅱ的最优混合策略(或简称最优策略)。

现约定，以下对 $G=\{S_1,S_2;\boldsymbol{A}\}$ 及其混合扩充 $G^*=\{S_1^*,S_2^*;E\}$ 一般不加区别，通常都用 $G=\{S_1,S_2;\boldsymbol{A}\}$ 来表示。当 G 在纯策略意义下解不存在时，自动认为讨论的是在混合策略意义下的解，相应的局中人Ⅰ的赢得函数为 $E(x,y)$。

和定理1类似，可给出矩阵对策 G 在混合策略意义下解存在的鞍点型充要条件。

定理 2 矩阵对策 $G=\{S_1,S_2;\boldsymbol{A}\}$ 在混合策略意义下有解的充要条件是：存在 $x^*\in S_1^*$，$y^*\in S_2^*$，使 (x^*,y^*) 为函数 $E(x,y)$ 的一个鞍点，即对一切 $x\in S_1^*$，$y\in S_2^*$，有

$$E(x,y^*)\leqslant E(x^*,y^*)\leqslant E(x^*,y) \tag{15-10}$$

本定理的证明同定理1，只需将 a_{ij} 换写成 $E(x,y)$ 读者可自行证之。应注意到，当 G 在纯策略意义下解存在时，定义4中关于对策 G 的值的定义 V_G 与前面的定义是一致的。当 G 在混合策略意义下的解 (x^*,y^*) 存在时，$V_G=E(x^*,y^*)$。

例 10 考虑矩阵对策 $G=\{S_1,S_2;\boldsymbol{A}\}$，其中

$$\boldsymbol{A}=\begin{bmatrix}3 & 6\\5 & 4\end{bmatrix}$$

由前面的讨论已知 G 在纯策略意义下的解不存在，于是设 $x=(x_1,x_2)$ 为局中人Ⅰ的混合策略，$y=(y_1,y_2)$ 为局中人Ⅱ的混合策略，则

$$S_1^*=\{(x_1,x_2)\mid x_1,x_2\geqslant 0,x_1+x_2=1\}$$
$$S_2^*=\{(y_1,y_2)\mid y_1,y_2\geqslant 0,y_1+y_2=1\}$$

局中人Ⅰ的赢得期望值是

$$\begin{aligned}E(x,y)&=3x_1y_1+6x_1y_2+5x_2y_1+4x_2y_2\\&=3x_1y_1+6x_1(1-y_1)+5y_1(1-x_1)+\\&\quad 4(1-x_1)(1-y_1)\\&=-4(x_1-1/4)(y_1-1/2)+9/2\end{aligned}$$

取 $\quad x^*=(1/4,3/4),y^*=(1/2,1/2)$

则 $\quad E(x^*,y^*)=9/2,E(x^*,y)=E(x,y^*)=9/2$

即有 $\quad E(x,y^*)\leqslant E(x^*,y^*)\leqslant E(x^*,y)$

故 $x^*=(1/4,3/4)$ 和 $y^*=(1/2,1/2)$ 分别为局中人Ⅰ和Ⅱ的最优策略，对策的值(局中人Ⅰ的赢得期望值)$V_G=9/2$。

15.2.3 矩阵对策的基本定理

本节主要讨论矩阵对策解的存在性及解的有关性质。如前所述，一般矩阵对策在纯策略意义下的解往往是不存在的。但本节将证明，一般矩阵对策在混合策略意义下的解却总是存在的，并且通过一个构造性的证明，引出求解矩阵对策的基本方法——线性规划方法。

先给出如下两个记号：

当局中人Ⅰ取纯策略 α_i 时，记其相应的赢得函数为 $E(i,y)$，于是

$$E(i,y)=\sum_j a_{ij}y_j \tag{15-11}$$

当局中人Ⅱ取纯策略 β_j 时，记其相应的赢得函数为 $E(x,j)$，于是

$$E(x,j)=\sum_i a_{ij}x_i \tag{15-12}$$

由式(15-11)和式(15-12)，有

$$\begin{aligned}E(x,y)&=\sum_i\sum_j a_{ij}x_iy_j=\sum_i\left(\sum_j a_{ij}y_j\right)x_i\\&=\sum_i E(i,y)x_i\end{aligned} \tag{15-13}$$

和

$$\begin{aligned}E(x,y)&=\sum_i\sum_j a_{ij}x_iy_j=\sum_j\left(\sum_i a_{ij}x_i\right)y_j\\&=\sum_j E(x,j)y_j\end{aligned} \tag{15-14}$$

根据上面的记号，可给出定理2的另一种等价表示。

定理 3 设 $x^* \in S_1^*, y^* \in S_2^*$，则 (x^*, y^*) 是 G 的解的充要条件是：对任意 $i=1,\cdots,m$ 和 $j=1,\cdots,n$，有

$$E(i,y^*) \leqslant E(x^*,y^*) \leqslant E(x^*,j) \tag{15-15}$$

证明 设 (x^*, y^*) 是 G 的解，则由定理 2，式(15-10)成立。由于纯策略是混合策略的特例，故式(15-15)成立。反之，设式(15-15)成立，由

$$E(x,y^*) = \sum_i E(i,y^*)x_i \leqslant E(x^*,y^*)\sum_i x_i = E(x^*,y^*)$$

$$E(x^*,y) = \sum_j E(x^*,j)y_j \geqslant E(x^*,y^*)\sum_j y_j = E(x^*,y^*)$$

即得式(15-10)。

定理 4 设 $x^* \in S_1^*, y^* \in S_2^*$，则 (x^*, y^*) 为 G 的解的充要条件是：存在数 v，使得 x^* 和 y^* 分别是不等式组(Ⅰ)和(Ⅱ)的解，且 $v=V_G$。

$$(\text{Ⅰ})\begin{cases} \sum_i a_{ij}x_i \geqslant v, & j=1,\cdots,n \\ \sum_i x_i = 1 & \\ x_i \geqslant 0, & i=1,\cdots,m \end{cases} \tag{15-16}$$

$$(\text{Ⅱ})\begin{cases} \sum_j a_{ij}y_j \leqslant v, & i=1,\cdots,m \\ \sum_j y_j = 1 & \\ y_j \geqslant 0, & j=1,\cdots,n \end{cases} \tag{15-17}$$

定理 5 对任一矩阵对策 $G=\{S_1,S_2;\boldsymbol{A}\}$，一定存在混合策略意义下的解。

证明 由定理 3，只要证明存在 $x^* \in S_1^*, y^* \in S_2^*$，使得式(15-15)成立。为此，考虑如下两个线性规划问题：

$$\max w$$

$$(\text{P})\begin{cases} \sum_i a_{ij}x_i \geqslant w, & j=1,\cdots,n \\ \sum_i x_i = 1 & \\ x_i \geqslant 0 & i=1,\cdots,m \end{cases}$$

和

$$\min v$$

$$(\text{D})\begin{cases} \sum_j a_{ij}y_j \leqslant v, & i=1,\cdots,m \\ \sum_j y_j = 1 & \\ y_j \geqslant 0 & j=1,\cdots,n \end{cases}$$

易验证，问题(P)和(D)是互为对偶的线性规划问题，而且

$$x=(1,0,\cdots,0)^{\mathrm{T}}\in E^m,\quad w=\min_j a_{1j}$$

是问题(P)的一个可行解。

$$y=(1,0,\cdots,0)^{\mathrm{T}}\in E^n,\quad v=\max_i a_{i1}$$

是问题(D)的一个可行解。由线性规划的对偶理论可知，问题(P)和(D)分别存在最优解(x^*,w^*)和(y^*,v^*)，且$v^*=w^*$。即存在$x^*\in S_1^*$，$y^*\in S_2^*$和数v^*，使得对任意$i=1,\cdots,m$和$j=1,\cdots,n$，有

$$\sum_j a_{ij}y_j^* \leqslant v^* \leqslant \sum_i a_{ij}x_i^* \tag{15-18}$$

或

$$E(i,y^*)\leqslant v^*\leqslant E(x^*,j) \tag{15-19}$$

又由

$$E(x^*,y^*)=\sum_i E(i,y^*)x_i^* \leqslant v^*\cdot\sum_i x_i^*=v^*$$

$$E(x^*,y^*)=\sum_j E(x^*,j)y_j^* \geqslant v^*\cdot\sum_j y_j^*=v^*$$

得到$v^*=E(x^*,y^*)$，故由式(15-19)知式(15-15)成立。 证毕。

定理5的证明是一个构造性的证明，不仅证明了矩阵对策解的存在性，而且给出了利用线性规划求解矩阵对策的方法。

下面的定理6至定理10进一步讨论了矩阵对策及其解的若干重要性质，这些性质在求解矩阵对策的过程中将起到重要作用。

定理6 设(x^*,y^*)是矩阵对策G的解，$v=V_G$，则

(1) 若$x_i^*>0$，则$\sum_j a_{ij}y_j^*=v$。

(2) 若$y_j^*>0$，则$\sum_i a_{ij}x_i^*=v$。

(3) 若$\sum_j a_{ij}y_j^*<v$，则$x_i^*=0$。

(4) 若$\sum_i a_{ij}x_i^*>v$，则$y_j^*=0$。

证明 按定义有

$$v=\max_{x\in S_1^*}E(x,y^*)$$

故

$$v-\sum_j a_{ij}y_j^*=\max_{x\in S_1^*}E(x,y^*)-E(i,y^*)\geqslant 0$$

又因

$$\sum_i x_i^*\left(v-\sum_j a_{ij}y_j^*\right)=v-\sum_i\sum_j a_{ij}x_i^*y_j^*=0$$

$$x_i^*\geqslant 0,\quad i=1,\cdots,m$$

所以，当 $x_i^*>0$ 时，必有 $\sum_j a_{ij}y_j^*=v$；当 $\sum_j a_{ij}y_j^*<v$ 时，必有 $x_i^*=0$，(1)、(3) 得证。同理可证(2)、(4)。

记矩阵对策 G 的解集为 $T(G)$，下面三个定理是关于对策解集性质的主要内容。

定理 7 设有两个矩阵对策

$$G_1=\{S_1,S_2;\boldsymbol{A}_1\}$$
$$G_2=\{S_1,S_2;\boldsymbol{A}_2\}$$

其中 $\boldsymbol{A}_1=(a_{ij})$，$\boldsymbol{A}_2=(a_{ij}+L)$，$L$ 为任一常数，则有

(1) $V_{G_2}=V_{G_1}+L$

(2) $T(G_1)=T(G_2)$

定理 8 设有两个矩阵对策

$$G_1=\{S_1,S_2;\boldsymbol{A}\}$$
$$G_2=\{S_1,S_2;\alpha\boldsymbol{A}\}$$

其中 $\alpha>0$ 为任一常数。则

(1) $V_{G_2}=\alpha V_{G_1}$

(2) $T(G_1)=T(G_2)$

定理 9 设 $G=\{S_1,S_2;\boldsymbol{A}\}$ 为一矩阵对策，且 $\boldsymbol{A}=-\boldsymbol{A}^{\mathrm{T}}$ 为斜对称矩阵(亦称这种对策为对称对策)。则

(1) $V_G=0$；

(2) $T_1(G)=T_2(G)$，其中 $T_1(G)$ 和 $T_2(G)$ 分别为局中人Ⅰ和Ⅱ的最优策略集。

定理 7 至定理 9 的证明留给读者去完成。在给出定理 10 之前，先给出矩阵对策优超纯策略的定义。

定义 5 设有矩阵对策 $G=\{S_1,S_2;\boldsymbol{A}\}$，其中 $S_1=\{\alpha_1,\cdots,\alpha_m\}$，$S_2=\{\beta_1,\cdots,\beta_n\}$，$\boldsymbol{A}=(a_{ij})$，如果对一切 $j=1,\cdots,n$ 都有 $a_{i^0j}\geqslant a_{k^0j}$，即矩阵 $\boldsymbol{A}$ 的第 i^0 行元素均不小于第 k^0 行的对应元素，则称局中人Ⅰ的纯策略 α_{i^0} 优超于 α_{k^0}；同样，若对一切 $i=1,\cdots,m$，都有 $a_{ij^0}\leqslant a_{il^0}$ 即矩阵 $\boldsymbol{A}$ 的第 l^0 列元素均不小于第 j^0 列的对应元素，则称局中人Ⅱ的纯策略 β_{j^0} 优超于 β_{l^0}。

定理 10 设 $G=\{S_1,S_2;\boldsymbol{A}\}$ 为矩阵对策，其中 $S_1=\{\alpha_1,\cdots,\alpha_m\}$，$S_2=\{\beta_1,\cdots,\beta_n\}$，$\boldsymbol{A}=(a_{ij})$ 如果纯策略 α_1 被其余纯策略 $\alpha_2,\cdots,\alpha_m$ 中之一所优超，由 G 可得到一个新的矩阵对策

$$G'=\{S_1',S_2;\boldsymbol{A}'\}$$

其中

$$S_1'=\{\alpha_2,\cdots,\alpha_m\}$$
$$\boldsymbol{A}'=(a_{ij}{}')_{(m-1)\times n}$$
$$a_{ij}'=a_{ij}\qquad i=2,\cdots,m;\ j=1,\cdots,n$$

于是有

(1) $V_{G'}=V_G$；

(2) G'中局中人Ⅱ的最优策略就是其在 G 中的最优策略；

(3) 若$(x_2^*,\cdots,x_m^*)^{\mathrm{T}}$ 是 G'中局中人Ⅰ的最优策略，则 $x^*=(0,x_2^*,\cdots,x_m^*)^{\mathrm{T}}$ 便是其在 G 中的最优策略。

证明 不妨设 α_2 优超于 α_1，即

$$a_{2j}\geqslant a_{1j},\quad j=1,\cdots,n \tag{15-20}$$

因 $x'^*=(x_2^*,\cdots,x_m^*)^{\mathrm{T}}$ 和 $y^*=(y_1^*,\cdots,y_n^*)$是 G'的解，由定理 3，有

$$\sum_{j=1}^{n}a_{ij}y_j^*\leqslant V_{G'}\leqslant\sum_{i=2}^{m}a_{ij}x_i^* \tag{15-21}$$
$$i=2,\cdots,m;\quad j=1,\cdots,n$$

因 α_2 优超于 α_1，由式(15-20)有

$$\sum_{j=1}^{n}a_{1j}y_i^*\leqslant\sum_{j=1}^{n}a_{2j}y_j^*\leqslant V_{G'} \tag{15-22}$$

合并式(15-21)和式(15-22)，得

$$\sum_{j=1}^{n}a_{ij}y_j^*\leqslant V_{G'}\leqslant\sum_{i=2}^{m}a_{ij}x_i^*+a_{1j}\cdot 0$$
$$i=1,\cdots,m;\quad j=1,\cdots,n$$

或

$$E(i,y^*)\leqslant V_{G'}\leqslant E(x^*,j)$$
$$i=1,\cdots,m;\quad j=1,\cdots,n$$

由定理 4，(x^*,y^*)是 G 的解，其中 $x^*=(0,x_2^*,\cdots,x_m^*)^{\mathrm{T}}$，且 $V_{G'}=V_G$。

推论 在定理 10 中，若 α_1 不是为纯策略$\alpha_2,\cdots,\alpha_m$ 中之一所优超，而是为 $\alpha_2,\cdots,\alpha_m$ 的某个凸线性组合所优超，定理的结论仍然成立。

定理 10 实际给出了一个化简赢得矩阵 $\boldsymbol{A}$ 的原则，称之为优超原则。根据这个原则，当局中人Ⅰ的某纯策略 α_i 被其他纯策略或纯策略的凸线性组合所优超时，可在矩阵 $\boldsymbol{A}$ 中划去第 i 行而得到一个与原对策 G 等价但赢得矩阵阶数较小的对策 G'，而 G'的求解往往比 G 的求解容易些，通过求解 G'而得到 G 的解。类似地，对局中人Ⅱ来说，可以在赢得矩阵 $\boldsymbol{A}$ 中划去被其他列或其他列的凸线性组合所优超的那些列。

下面，举例说明优超原则的应用。

例 11 设赢得矩阵为 $\boldsymbol{A}$，求解这个矩阵对策。

$$\boldsymbol{A}=\begin{bmatrix}3 & 2 & 0 & 3 & 0\\ 5 & 0 & 2 & 5 & 9\\ 7 & 3 & 9 & 5 & 9\\ 4 & 6 & 8 & 7 & 5.5\\ 6 & 0 & 8 & 8 & 3\end{bmatrix}$$

解 由于第 4 行优超于第 1 行，第 3 行优超于第 2 行，故可划去第 1 行和第 2 行，得

到新的赢得矩阵

$$A_1=\begin{bmatrix}7 & 3 & 9 & 5 & 9\\4 & 6 & 8 & 7 & 5.5\\6 & 0 & 8 & 8 & 3\end{bmatrix}$$

对于 A_1，第1列优超于第3列，第2列优超于第4列，1/3×(第1列)+2/3×(第2列)优超于第5列，因此去掉第3列、第4列和第5列，得到

$$A_2=\begin{bmatrix}7 & 3\\4 & 6\\6 & 0\end{bmatrix}$$

这时，第1行又优超于第3行，故从 A_2 中划去第3行，得到

$$A_3=\begin{bmatrix}7 & 3\\4 & 6\end{bmatrix}$$

对于 A_3，易知无鞍点存在，应用定理4，求解不等式组(Ⅰ)，不等式组(Ⅱ)

$$(\text{Ⅰ})\begin{cases}7x_3+4x_4\geqslant v\\3x_3+6x_4\geqslant v\\x_3+x_4=1\\x_3,x_4\geqslant 0\end{cases}\qquad(\text{Ⅱ})\begin{cases}7y_1+3y_2\leqslant v\\4y_1+6y_2\leqslant v\\y_1+y_2=1\\y_1,y_2\geqslant 0\end{cases}$$

首先考虑满足

$$\begin{cases}7x_3+4x_4=v\\3x_3+6x_4=v\\x_3+x_4=1\end{cases}\qquad\begin{cases}7y_1+3y_2=v\\4y_1+6y_2=v\\y_1+y_2=1\end{cases}$$

的非负解。求得解为

$$x_3^*=1/3,\quad x_4^*=2/3$$
$$y_1^*=1/2,\quad y_2^*=1/2$$
$$v=5$$

于是，原矩阵对策的一个解为

$$x^*=(0,0,1/3,2/3,0)^{\mathrm{T}}$$
$$y^*=(1/2,1/2,0,0,0)^{\mathrm{T}}$$
$$V_G=5$$

15.3 矩阵对策的解法

15.3.1 公式法、图解法和方程组法

1. 2×2对策的公式法

所谓2×2对策是指局中人Ⅰ的赢得矩阵为2×2阶的，即

$$A=\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

如果 A 有鞍点，则很快可求出各局中人的最优纯策略；如果 A 没有鞍点，则可证明各局中人最优混合策略中的 x_i^*，y_j^* 均大于零。于是，由定理 6 可知，为求最优混合策略可求下列等式组：

$$(\text{I})\begin{cases} a_{11}x_1+a_{21}x_2=v \\ a_{12}x_1+a_{22}x_2=v \\ x_1+\quad x_2=1 \end{cases} \tag{15-23}$$

$$(\text{II})\begin{cases} a_{11}y_1+a_{12}y_2=v \\ a_{21}y_1+a_{22}y_2=v \\ y_1+\quad y_2=1 \end{cases} \tag{15-24}$$

当矩阵 A 不存在鞍点时，可以证明上面等式组(Ⅰ)和等式组(Ⅱ)一定有严格非负解 $x^*=(x_1^*,x_2^*)$和 $y^*=(y_1^*,y_2^*)$，其中

$$x_1^*=\frac{a_{22}-a_{21}}{(a_{11}+a_{22})-(a_{12}+a_{21})} \tag{15-25}$$

$$x_2^*=\frac{a_{11}-a_{12}}{(a_{11}+a_{22})-(a_{12}+a_{21})} \tag{15-26}$$

$$y_1^*=\frac{a_{22}-a_{12}}{(a_{11}+a_{22})-(a_{12}+a_{21})} \tag{15-27}$$

$$y_2^*=\frac{a_{11}-a_{21}}{(a_{11}+a_{22})-(a_{12}+a_{21})} \tag{15-28}$$

$$V_G=\frac{a_{11}a_{22}-a_{12}a_{21}}{(a_{11}+a_{22})-(a_{12}+a_{21})} \tag{15-29}$$

例 12 求解矩阵对策 $G=\{S_1,S_2;A\}$，其中

$$A=\begin{bmatrix} 1 & 3 \\ 4 & 2 \end{bmatrix}$$

解 易知，A 没有鞍点。由通解式(15-25)～式(15-29)计算得到最优解为 $x^*=(1/2,1/2)^T$，$y^*=(1/4,3/4)^T$，对策值为 5/2。

2. 2×n 或 m×2 对策的图解法

这里，介绍一种求矩阵对策的图解法，这个方法用在赢得矩阵为 $2\times n$ 或 $m\times 2$ 阶的对策上特别方便，也可用在 $3\times n$ 或 $m\times 3$ 对策上。但对 m 和 n 均大于 3 的矩阵对策就不适用了。现用一个 $2\times n$ 对策的例子来说明这个方法。

例 13 考虑矩阵对策 $G=\{S_1,S_2;A\}$，其中

$$A=\begin{bmatrix} 2 & 3 & 11 \\ 7 & 5 & 2 \end{bmatrix}$$

$$S_1=\{\alpha_1,\alpha_2\},\quad S_2=\{\beta_1,\beta_2,\beta_3\}$$

设局中人Ⅰ的混合策略为$(x,1-x)^T$，$x\in[0,1]$。过数轴上坐标为O和$(1,0)$的两点分别作两条垂线Ⅰ－Ⅰ和Ⅱ－Ⅱ，垂线上点的纵坐标值分别表示局中人Ⅰ采取纯策略α_1和α_2时，局中人Ⅱ采取各纯策略时的赢得值。如图15-1所示。当局中人Ⅰ选择每一策略$(x,1-x)^T$时，他的最少可能的收入为由局中人Ⅱ选择β_1、β_2、β_3时所确定的三条直线

$$2x+7(1-x)=V$$
$$3x+5(1-x)=V$$
$$11x+2(1-x)=V$$

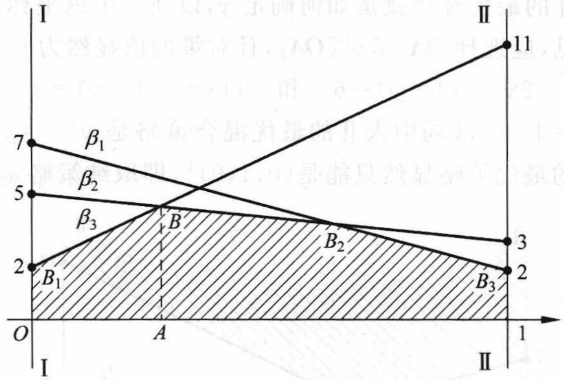


图15-1　$2\times n$对策的图解法

在x处的纵坐标中之最小者，即如折线$B_1BB_2B_3$所示。所以对局中人Ⅰ来说，他的最优选择就是确定x使他的收入尽可能地多，从图15-1可知，按最小最大原则应选择$x=OA$，而AB即为对策值。为求出点x和对策值V_G，可联立过B点的两条线段β_2和β_3所确定的方程：

$$\begin{cases}3x+5(1-x)=V_G\\11x+2(1-x)=V_G\end{cases}$$

解得$x=3/11$，$V_G=49/11$。所以，局中人Ⅰ的最优策略为$x^*=(3/11,8/11)^T$。此外，从图15-1还可以看出，局中人Ⅱ的最优混合策略只由β_2和β_3组成。

事实上，若记$y^*=(y_1^*,y_2^*,y_3^*)^T$为局中人Ⅱ的最优混合策略，则由

$$E(x^*,1)=2\times3/11+7\times8/11=62/11>49/11=V_G$$
$$E(x^*,2)=E(x^*,3)=V_G$$

根据定理6可知，必有$y_1^*=0$，$y_2^*>0$，$y_3^*>0$。

根据定理6，可由

$$\begin{cases}3y_2+11y_3=49/11\\5y_2+\ \ 2y_3=49/11\\\ \ y_2+\ \ \ \ y_3=1\end{cases}$$

求得 $y_2^*=9/11, y_3^*=2/11$。所以局中人Ⅱ的最优混合策略为 $y^*=(0,9/11,2/11)^T$。

例 14 用图解法求解矩阵对策 $G=\{S_1,S_2,\boldsymbol{A}\}$，其中 $S_1=\{\alpha_1,\alpha_2,\alpha_3\}$，$S_2=\{\beta_1,\beta_2\}$，

$$\boldsymbol{A}=\begin{bmatrix}2 & 7\\ 6 & 6\\ 11 & 2\end{bmatrix}$$

设局中人Ⅱ的混合策略为 $(y,1-y)^T$，由图 15-2 可知，直线 α_1、α_2、α_3 在任一点 $y\in[0,1]$ 处的纵坐标分别是局中人Ⅱ采取混合策略 $(y,1-y)^T$ 时的支付。根据最不利当中选取最有利的原则，局中人Ⅱ的最优选择就是如何确定 y，以使三个纵坐标值中的最大值尽可能地小。从图 15-2 可见，应选择 $\boldsymbol{OA}_1\leqslant y\leqslant \boldsymbol{OA}_2$，且对策的值显然为 6。由方程

$$2y+7(1-y)=6 \quad 和 \quad 11y+2(1-y)=6$$

求得 $\boldsymbol{OA}_1=1/5$，$\boldsymbol{OA}_2=4/9$。故局中人Ⅱ的最优混合策略是 $y^*=(y,1-y)^T$，其中 $1/5\leqslant y\leqslant 4/9$，而局中人Ⅰ的最优策略显然只能是 $(0,1,0)^T$，即取纯策略 α_2。

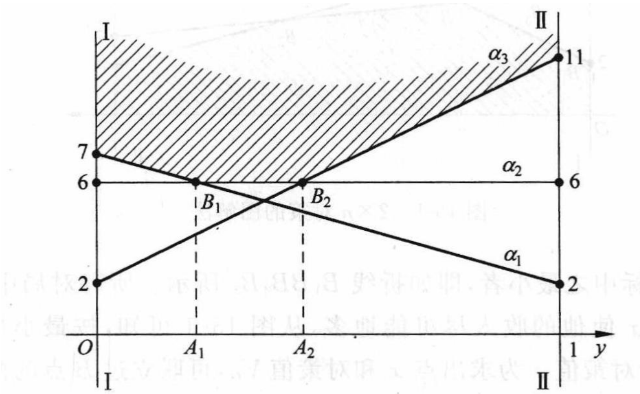


图 15-2 $m\times 2$ 对策的图解法

例 15 求解赢得矩阵 $\boldsymbol{A}$ 的矩阵对策。

$$\boldsymbol{A}=\begin{bmatrix}4 & 8/3 & 4 & 2\\ 1 & 5 & 5 & 7\end{bmatrix}$$

首先，利用优超原则，第 2 列优超于第 3 列，故可划去第 3 列，又因

$$(2/3)\times(第4列)+(1/3)\times(第1列)=(第2列)$$

由优超原则又可划去 $\boldsymbol{A}$ 的第 2 列，转而求解赢得矩阵

$$\boldsymbol{A}'=\begin{bmatrix}4 & 2\\ 1 & 7\end{bmatrix}$$

的对策 G'，从而解得原对策的一个解为

$$x^* = (3/4, 1/4)^T$$
$$y^* = (5/8, 0, 0, 3/8)^T$$
$$V_G = 13/4$$

若用图解法求解本例，由图 15-3 可见，局中人 Ⅰ 的最优策略为 $x^* = (3/4, 1/4)^T$，若记局中人 Ⅱ 的最优策略为 $y^* = (y_1^*, y_2^*, y_3^*, y_4^*)^T$，则有 $y_3^* = 0$，而 y_1^*, y_2^*, y_4^* 可由以下联立方程确定。

$$\begin{cases} 4y_1 + (8/3)y_2 + 2y_4 = 13/4 \\ y_1 + 5y_2 + 7y_4 = 13/4 \\ y_1 + y_2 + y_4 = 1 \end{cases}$$

显然，满足上面方程组的解有无穷多个，故局中人Ⅱ有无穷多个最优混合策略。

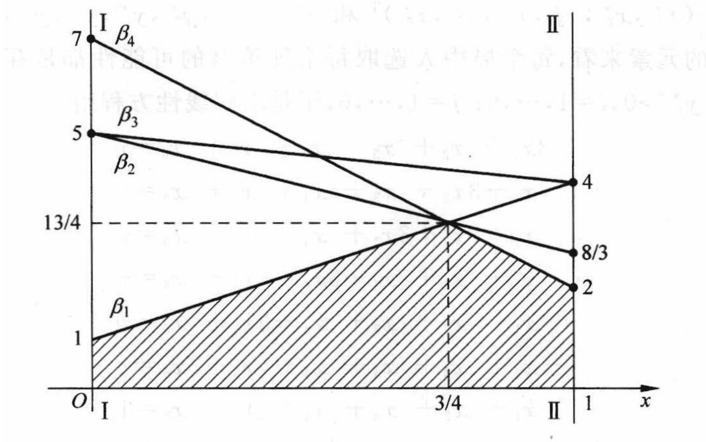


图 15-3

例 15 说明，利用优超原则化简赢得矩阵时，有可能将原矩阵对策的解也划去一些，这种情况在 m 和 n 均大于 3 时仍然可能发生。

3. 线性方程组方法

根据定理 4，求解矩阵对策解 (x^*, y^*) 的问题等价于求解不等式组(15-16)和(15-17)，又根据定理 5 和定理 6，如果假设最优策略中的 x_i^* 和 y_j^* 均不为零，即可将上述两个不等式组的求解问题转化成求解下面两个方程组的问题

$$\begin{cases} \sum_i a_{ij} x_i = v, \quad j = 1, \cdots, n \\ \sum_i x_i = 1 \end{cases} \tag{15-30}$$

和

$$\begin{cases} \sum_j a_{ij} y_j = v, \quad i = 1, \cdots, m \\ \sum_j y_j = 1 \end{cases} \tag{15-31}$$

例 16 求解矩阵对策——“齐王赛马”。

解 已知齐王的赢得矩阵为

$$A=\begin{bmatrix} 3 & 1 & 1 & 1 & 1 & -1 \\ 1 & 3 & 1 & 1 & -1 & 1 \\ 1 & -1 & 3 & 1 & 1 & 1 \\ -1 & 1 & 1 & 3 & 1 & 1 \\ 1 & 1 & -1 & 1 & 3 & 1 \\ 1 & 1 & 1 & -1 & 1 & 3 \end{bmatrix}$$

易知，A 没有鞍点，即对齐王和田忌来说都不存在最优纯策略。设齐王和田忌的最优混合策略为

$$x^*=(x_1^*,x_2^*,x_3^*,x_4^*,x_5^*,x_6^*)^T \text{ 和 } y^*=(y_1^*,y_2^*,y_3^*,y_4^*,y_5^*,y_6^*)^T$$

从矩阵 A 的元素来看，每个局中人选取每个纯策略的可能性都是存在的，故可事先假定 $x_i^*>0$ 和 $y_j^*>0$，$i=1,\cdots,6$；$j=1,\cdots,6$，于是求解线性方程组

$$\begin{cases} 3x_1+x_2+x_3-x_4+x_5+x_6=v \\ x_1+3x_2-x_3+x_4+x_5+x_6=v \\ x_1+x_2+3x_3+x_4-x_5+x_6=v \\ x_1+x_2+x_3+3x_4+x_5-x_6=v \\ x_1-x_2+x_3+x_4+3x_5+x_6=v \\ -x_1+x_2+x_3+x_4+x_5+3x_6=v \\ x_1+x_2+x_3+x_4+x_5+x_6=1 \end{cases} \tag{15-32}$$

和

$$\begin{cases} 3y_1+y_2+y_3+y_4+y_5-y_6=v \\ y_1+3y_2+y_3+y_4-y_5+y_6=v \\ y_1-y_2+3y_3+y_4+y_5+y_6=v \\ -y_1+y_2+y_3+3y_4+y_5+y_6=v \\ y_1+y_2-y_3+y_4+3y_5+y_6=v \\ y_1+y_2+y_3-y_4+y_5+3y_6=v \\ y_1+y_2+y_3+y_4+y_5+y_6=1 \end{cases} \tag{15-33}$$

得到

$$x_i=1/6,\quad i=1,\cdots,6$$
$$y_j=1/6,\quad j=1,\cdots,6$$
$$v=1$$

故齐王和田忌的最优混合策略为

$$x^*=(1/6,1/6,1/6,1/6,1/6,1/6)^T \text{ 和 } y^*=(1/6,1/6,1/6,1/6,1/6,1/6)^T$$

对策的值(齐王的期望赢得)为 $V_G=1$。这与我们的设想相符,即双方都以 1/6 的概率选取每个纯策略,或者说每个纯策略被选取的机会应是均等的,则总的结局应该是:齐王有 5/6 的机会赢田忌,赢得的期望值是 1 千金。但是,如果齐王在每出一匹马前将自己的选择告诉了对方,这实际上等于公开了自己的策略,如齐王选取出马次序为(上,中,下),则田忌根据谋士的建议便以(下,上,中)对之,结果田忌反而可得千金。因此,在矩阵对策不存在鞍点时,竞争的双方在开局前均应对自己的策略(实际上是纯策略)加以保密,否则不保密的一方是要吃亏的。

例 17 某厂用三种不同的设备 α_1、α_2、α_3 加工三种不同的产品 β_1、β_2、β_3,已知三种设备分别加工三种产品时,单位时间内创造的价值由表 15-4 给出。

表 15-4

使用设备	被加工产品		
	β_1	β_2	β_3
α_1	3	−2	4
α_2	−1	4	2
α_3	2	2	6

出现负值是由于设备的消耗大于创造出的价值。在上述条件下,求出一个合理的加工方案。

解 此问题可看成是一个矩阵对策问题,并易知没有鞍点。设采用设备 α_1、α_2、α_3 的概率分别为 $(x_1,x_2,x_3)^{\mathrm{T}}$,产品 β_1、β_2、β_3 被接受加工的概率分别为 $(y_1,y_2,y_3)^{\mathrm{T}}$。

为简化求解计算,由定理 7,转而求赢得矩阵为

$$\boldsymbol{A}'=\boldsymbol{A}-(2)_{3\times3}=\begin{bmatrix}1 & -4 & 2\\ -3 & 2 & 0\\ 0 & 0 & 4\end{bmatrix}$$

的矩阵对策 G',为此,先求解等式组

$$\begin{cases} x_1-3x_2 = v' \\ -4x_1+2x_2 = v' \\ 2x_1+4x_3=v' \\ x_1+x_2+x_3=1 \end{cases} \tag{15-34}$$

和

$$\begin{cases} y_1-4y_2+2y_3=v' \\ -3y_1+2y_2 = v' \\ 4y_3=v' \\ y_1+y_2+y_3=1 \end{cases} \tag{15-35}$$

经过实际计算,知道上述等式组不存在非负解。于是根据试算,将式(15-34)中第 3 式取

为不等式,将式(15-35)中第 1 式取为不等式,转而求解

$$\begin{cases} x_1 - 3x_2 = v' \\ -4x_1 + 2x_2 = v' \\ 2x_1 + 4x_3 > v' \\ x_1 + x_2 + x_3 = 1 \end{cases} \tag{15-36}$$

和

$$\begin{cases} y_1 - 4y_2 + 2y_3 < v' \\ -3y_1 + 2y_2 = v' \\ 4y_3 = v' \\ y_1 + y_2 + y_3 = 1 \end{cases} \tag{15-37}$$

由定理 6,式(15-36)和式(15-37)的解 x^* 及 y^* 中的分量 $x_1^*=0, y_3^*=0$。将此结果代回上面两组不等式,得到

$$x_2^*=0, x_3^*=1$$
$$y_1^*=2/5,\quad y_2^*=3/5$$
$$v'=0$$

故原对策最优解为 $x^*=(0,0,1)^{\mathrm{T}}$, $y^*=(2/5,3/5,0)^{\mathrm{T}}$。

对策值(单位时间内创造价值的期望值)$V_G=v'+2=2$。

15.3.2 线性规划方法

由定理 5 已知,任一矩阵对策 $G=\{S_1,S_2;\boldsymbol{A}\}$的求解均等价于一对互为对偶的线性规划问题,而定理 4 表明,对策 G 的解 x^* 和 y^* 等价于下面两个不等式组的解。

$$(\mathrm{I})\begin{cases} \sum_i a_{ij}x_i \geqslant v, \quad j=1,\cdots,n \\ \sum_i x_i = 1 \\ x_i \geqslant 0, \quad i=1,\cdots,m \end{cases} \tag{15-38}$$

$$(\mathrm{II})\begin{cases} \sum_j a_{ij}y_j \leqslant v, \quad i=1,\cdots,m \\ \sum_j y_j = 1 \\ y_j \geqslant 0, \quad j=1,\cdots,n \end{cases} \tag{15-39}$$

其中

$$v=\max_{x\in S_1^*}\min_{y\in S_2^*} E(x,y)=\min_{y\in S_2^*}\max_{x\in S_1^*} E(x,y) \tag{15-40}$$

就是对策的值 V_G。

定理 11 设矩阵对策 $G=\{S_1,S_2;\boldsymbol{A}\}$ 的值为 V_G，则

$$V_G=\max_{x\in S_1^*}\min_{1\leqslant j\leqslant n}E(x,j)=\min_{y\in S_2^*}\max_{1\leqslant i\leqslant m}E(i,y) \tag{15-41}$$

证明 因 V_G 是对策的值，故

$$V_G=\max_{x\in S_1^*}\min_{y\in S_2^*}E(x,y)=\min_{y\in S_2^*}\max_{x\in S_1^*}E(x,y) \tag{15-42}$$

一方面，任给 $x\in S_1^*$，有

$$\min_{1\leqslant j\leqslant n}E(x,j)\geqslant\min_{y\in S_2^*}E(x,y) \tag{15-43}$$

故

$$\max_{x\in S_1^*}\min_{1\leqslant j\leqslant n}E(x,j)\geqslant\max_{x\in S_1^*}\min_{y\in S_2^*}E(x,y) \tag{15-44}$$

另一方面，任给 $x\in S_1^*$，$y\in S_2^*$，有

$$E(x,y)=\sum_{j=1}^{n}E(x,j)\cdot y_j\geqslant\min_{1\leqslant j\leqslant n}E(x,j)$$

故

$$\min_{y\in S_2^*}E(x,y)\geqslant\min_{1\leqslant j\leqslant n}E(x,j) \tag{15-45}$$

$$\max_{x\in S_1^*}\min_{y\in S_2^*}E(x,y)\geqslant\max_{x\in S_1^*}\min_{1\leqslant j\leqslant n}E(x,j) \tag{15-46}$$

由式(15-44)和式(15-46)得

$$V_G=\max_{x\in S_1^*}\min_{1\leqslant j\leqslant n}E(x,j)$$

同理可证

$$V_G=\min_{y\in S_2^*}\max_{1\leqslant i\leqslant m}E(i,y)$$

证毕。

下面给出求解矩阵对策的线性规划方法。

作变换(根据定理 7，不妨设 $v>0$)

$$x_i'=x_i/v,\quad i=1,\cdots,m \tag{15-47}$$

则不等式组(15-38)变为

$$(\mathrm{I})\begin{cases}\sum_i a_{ij}x'_i\geqslant 1, & j=1,\cdots,n\\ \sum_i x'_i=1/v\\ x'_i\geqslant 0, & i=1,\cdots,m\end{cases} \tag{15-48}$$

根据定理 11，$v=\max_{x\in S_1^*}\min_{1\leqslant j\leqslant n}\sum_i a_{ij}x_i$。

这样，不等式组(Ⅰ)即等价于线性规划问题

$$
(\mathrm{P})\begin{cases}\min z=\sum_i x'_i \\ \sum_i a_{ij}x'_i\geqslant 1, \quad j=1,\cdots,n \\ x'_i\geqslant 0\end{cases} \tag{15-49}
$$

同理,作变换

$$
y'_j=y_j/v, \quad j=1,\cdots,n \tag{15-50}
$$

则不等式组(15-39)变为

$$
(\mathrm{II})\begin{cases}\sum_j a_{ij}y'_j\leqslant 1, \quad i=1,\cdots,m \\ \sum_j y'_j=1/v \\ y'_j\geqslant 0, \quad j=1,\cdots,n\end{cases} \tag{15-51}
$$

其中 $v=\min\limits_{y\in S_2^*}\max\limits_{1\leqslant i\leqslant m}\sum_j a_{ij}y_j$,与之等价的线性规划问题是

$$
(\mathrm{D})\begin{cases}\max w=\sum_j y'_j \\ \sum_j a_{ij}y'_j\leqslant 1, \quad i=1,\cdots,m \\ y'_j\geqslant 0, \quad j=1,\cdots,n\end{cases} \tag{15-52}
$$

显然,问题(P)和问题(D)是互为对偶的线性规划,故可利用单纯形或对偶单纯形方法求解。在求解时,一般先求问题(D)的解,因为这样容易在迭代的第一步就找到第一个基本可行解,而问题(P)的解从问题(D)的最后一个单纯形表上即可得到。当求得问题(P)和问题(D)的解后,再利用变换式(15-47)和式(15-50)即可求出原对策问题的解及对策的值。

例 18 利用线性规划方法求解赢得矩阵为 $\boldsymbol{A}$ 的矩阵对策。

$$
\boldsymbol{A}=\begin{bmatrix}7 & 2 & 9\\ 2 & 9 & 0\\ 9 & 0 & 11\end{bmatrix}
$$

解 求解问题可化成两个互为对偶的线性规划问题

$$
\min(x_1+x_2+x_3)
$$

$$
\begin{cases}7x_1+2x_2+9x_3\geqslant 1 \\ 2x_1+9x_2\geqslant 1 \\ 9x_1+11x_3\geqslant 1 \\ x_1,x_2,x_3\geqslant 0\end{cases} \tag{15-53}
$$

$$\max(y_1+y_2+y_3)$$
$$\begin{cases}7y_1+2y_2+9y_3\leqslant 1\\2y_1+9y_2\leqslant 1\\9y_1+11y_3\leqslant 1\\y_1,y_2,y_3\geqslant 0\end{cases}\tag{15-54}$$

利用单纯形方法求解问题(D)，迭代过程如表 15-5 所示。

表 15-5

c_j			1	1	1	0	0	0	w
C_B	X_B	b	y_1	y_2	y_3	u_1	u_2	u_3	
0	u_1	1	7	2	9	1	0	0	
0	u_2	1	2	9	0	0	1	0	
0	u_3	1	[9]	0	11	0	0	1	
c_j-z_j			1	1	1	0	0	0	0
0	u_1	2/9	0	2	4/9	1	0	−7/9	
0	u_2	7/9	0	[9]	−22/9	0	1	−2/9	
1	y_1	1/9	1	0	11/9	0	0	1/9	
c_j-z_j			0	1	−2/9	0	0	−1/9	1/9
0	u_1	4/81	0	0	[80/81]	1	−2/9	−59/81	
1	y_2	7/81	0	1	−22/81	0	1/9	−2/81	
1	y_1	1/9	1	0	11/9	0	0	1/9	
c_j-z_j			0	0	4/81	0	−1/9	−7/81	16/81
1	y_3	4/80	0	0	1	81/80	−18/80	−59/80	
1	y_2	1/10	0	1	0	22/80	4/80	−18/80	
1	y_1	1/20	1	0	0	−99/80	22/80	81/80	
c_j-z_j			0	0	0	−4/80	−8/80	−4/80	16/80

从表 15-5 中可得到问题(D)的解为

$$\begin{cases}y=(1/20,1/10,4/80)^{\mathrm{T}}=(1/20,1/10,1/20)^{\mathrm{T}}\\w=16/80\end{cases}$$

由表 15-5 中最后一个单纯形表可得问题(P)的解为

$$\begin{cases}x=(4/80,8/80,4/80)^{\mathrm{T}}=(1/20,1/10,1/20)^{\mathrm{T}}\\z=16/80\end{cases}$$

于是

$$V_G=80/16=5$$

$$x^* = V_G \cdot (1/20,1/10,1/20)^T = (1/4,2/4,1/4)^T$$
$$y^* = V_G \cdot (1/20,1/10,1/20)^T = (1/4,2/4,1/4)^T$$

至此，我们介绍了一些求解矩阵对策的方法。在求解一个矩阵对策时，应首先判断其是否具有鞍点，当鞍点不存在时，利用优超原则和定理7、定理8等提供的方法将原对策的赢得矩阵尽量地化简，然后再利用本节介绍的各种方法去求解。

在本节介绍的各类求解的方法中，线性规划方法是具有一般性的，另外还有两种具有一般性的解法：求全部解的矩阵法和至少保证求出一个解的微分方程法。限于篇幅，这里就不作介绍了，有兴趣的读者可参阅参考资料[1]。

15.4* 其他类型对策简介

15.4.1 二人无限零和对策

矩阵对策最简单的推广就是局中人的策略集从有限集变为无限集，例如是[0,1]区间。一般用 $G=\{S_1,S_2;H\}$ 表示一个二人无限零和对策，其中 S_1 和 S_2 中至少有一个是无限集合，H 为局中人Ⅰ的赢得函数。记

$$v_1 = \max_{\alpha_i \in S_1} \min_{\beta_j \in S_2} H(\alpha_i,\beta_j)$$

$$v_2 = \min_{\beta_j \in S_2} \max_{\alpha_i \in S_1} H(\alpha_i,\beta_j)$$

则 v_1 为局中人Ⅰ的至少赢得，v_2 为局中人Ⅱ的至多所失。显然有 $v_1 \leqslant v_2$，当 $v_1 = v_2$ 时，有如下定义。

定义 6 设 $G=\{S_1,S_1;H\}$ 为二人无限零和对策。若存在 $\alpha_{i^*} \in S_1, \beta_{j^*} \in S_2$，使得

$$\max_{\alpha_i \in S_1} \min_{\beta_j \in S_2} H(\alpha_i,\beta_j) = \min_{\beta_j \in S_2} \max_{\alpha_i \in S_1} H(\alpha_i,\beta_j) = H(\alpha_{i^*},\beta_{j^*}) \quad (15\text{-}55)$$

记其值为 V_G，则称 V_G 为对策 G 的值，称使式(15-55)成立的 $(\alpha_{i^*},\beta_{j^*})$ 为 G 在纯策略意义下的解，α_{i^*},β_{j^*} 分别称为局中人Ⅰ和Ⅱ的最优纯策略。

定理 12 $(\alpha_{i^*},\beta_{j^*})$ 为 $G=\{S_1,S_2;H\}$ 在纯策略意义下的解的充要条件是：对任意 $\alpha_i \in S_1, \beta_j \in S_2$，有

$$H(\alpha_i,\beta_{j^*}) \leqslant H(\alpha_{i^*},\beta_{j^*}) \leqslant H(\alpha_{i^*},\beta_j) \quad (15\text{-}56)$$

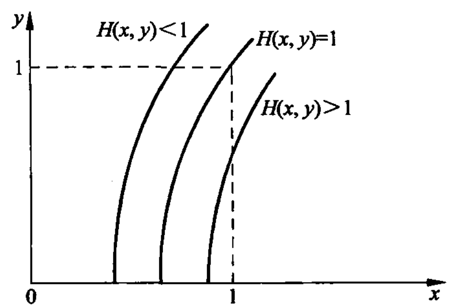


图 15-4

例 19 设局中人Ⅰ、Ⅱ互相独立地从[0,1]中分别选择一个实数 x 和 y，局中人Ⅰ的赢得函数为 $H(x,y)=2x^2-y^2$。对策中，局中人Ⅰ希望 H 越大越好，局中人Ⅱ则希望 H 越小越好。图15-4给出了 $H(x,y)$ 的等值线，通过对该图的分析，不难看出双方竞争的平衡局势为(1,1)，

即 $\alpha_{i^*}=1,\beta_{j^*}=1$ 分别为局中人Ⅰ和Ⅱ的最优纯策略，$V_G=1$。可以验证，对$(\alpha_{i^*},\beta_{j^*})=(1,1)$，式(15-56)是成立的。

由矩阵对策的结果已知，(15-55)式一般并不成立，即对策 $G=\{S_1,S_2;H\}$ 在纯策略意义下无解。同矩阵对策中引入混合策略的做法类似，也可定义无限对策的混合策略如下：局中人Ⅰ和Ⅱ的混合策略 X 和 Y 分别为策略集 S_1 和 S_2 上的概率分布(或分布函数)，混合策略集记为 $\bar{X}$ 和 $\bar{Y}$。若用于 x,y 表示纯策略，$F_X(x)$，$F_Y(y)$表示混合策略 X,Y 的分布，则局中人Ⅰ的赢得函数可以有以下4种形式

$$H(x,y)$$

$$H(X,y)=\int_{S_1}H(x,y)\mathrm{d}F_X(x)$$

$$H(x,Y)=\int_{S_2}H(x,y)\mathrm{d}F_Y(y)$$

$$H(X,Y)=\int_{S_1}\int_{S_2}H(x,y)\mathrm{d}F_X(x)F_Y(y)$$

定义7　如果有

$$\sup_X\inf_Y H(X,Y)=\inf_Y\sup_X H(X,Y)=V_G \tag{15-57}$$

则称 V_G 为对策 G 的值，称使式(15-57)成立的(X^*,Y^*)为对策 G 的解，X^* 和 Y^* 分别为局中人Ⅰ和Ⅱ的最优策略。

定理13　(X^*,Y^*)为对策 $G=\{S_1,S_2;H\}$的解的充要条件是：对任意 $X\in\bar{X},Y\in\bar{Y}$，有

$$H(X,Y^*)\leqslant H(X^*,Y^*)\leqslant H(X^*,Y) \tag{15-58}$$

当 $S_1=S_2=[0,1]$，且 $H(x,y)$为连续函数时，称这样的对策为连续对策。对连续对策而言，局中人Ⅰ、Ⅱ的混合策略即为[0,1]区间上的分布函数。记[0,1]区间上的分布函数的集合为 D，则有

$$H(X,Y)=\int_0^1\int_0^1 H(x,y)\mathrm{d}F_X(x)\mathrm{d}F_Y(y)$$

对连续对策，记

$$v_1=\max_{X\in D}\min_{Y\in D}H(X,Y)$$

$$v_2=\min_{Y\in D}\max_{X\in D}H(X,Y)$$

这里，不加证明地给出关于连续对策的基本定理。

定理14　对任何连续对策，一定有 $v_1=v_2$。

例20　(生产能力分配问题)　某公司下属甲、乙两个工厂，分别位于 A、B 两市。设两厂总生产能力为1个单位，两市对工厂产品的总需求也是1个单位。如果 A 市的需求量为 x，则 B 市的需求量为 $1-x$，这时只要安排 A 厂的生产能力为 x，就能使供需平衡。但现在不知道 A 市的确切需求量 x 是多少，如果安排 A 厂的生产能力为 y，则将产生供

需上的不平衡。不平衡的程度可用数值表示为

$$\max\left\{\frac{x}{y},\frac{1-x}{1-y}\right\}$$

公司的目标是选择 y,使得

$$\max_{0\leqslant x\leqslant 1}\max\left\{\frac{x}{y},\frac{1-x}{1-y}\right\} \tag{15-59}$$

达到极小。如果以市场需求为一方,公司为另一方,则以上问题可转化为一个连续对策问题,其中

$$S_1=S_2=[0,1],H(x,y)=\max\left\{\frac{x}{y},\frac{1-x}{1-y}\right\}$$

对这个对策求解的结果为:公司方的最优策略(为纯策略)是 $y^*=1/2$,即两个厂各生产一半;市场需求方的最优策略(为混合策略)是:分别以 0.5 的概率取 0 和 1,即要么全部需求都集中在 A 市,要么都集中在 B 市,且两种情况发生的概率相等。该对策的值为 $V_G=2$,即当公司和市场均选择各自的最优策略时,两市中需求大于供给的平均程度为 2(求解过程较复杂,略去)。

15.4.2 多人非合作对策

实际问题中,会经常出现多人对策的问题,且每个局中人的赢得函数之和也不必一定为零,特别是许多经济过程中的对策模型一般都是非零和的,因为经济过程总是有新价值的产生。所谓非合作对策,就是指局中人之间互不合作,对策略的选择不允许事先有任何交换信息的行为,不允许订立任何约定,矩阵对策就是一种非合作对策。一般非合作对策模型可描述为:

(1) 局中人集合:$I=\{1,2,\cdots,n\}$;

(2) 每个局中人的策略集:$S_1,S_2,\cdots,S_n$(均为有限集);

(3) 局势:$s=(s_1,\cdots,s_n)\in S_1\times\cdots\times S_n$;

(4) 每个局中人 i 的赢得函数记为 $H_i(s)$,一般说来,$\sum_{i=1}^{n}H_i(s)\neq 0$。一个非合作 n 人对策一般用符号 $G=\{I,\{S_i\},\{H_i\}\}$ 表示。

为讨论非合作 n 人对策的平衡局势,引入记号

$$s\parallel s_i^0=(s_1,\cdots,s_{i-1},s_i^0,s_{i+1},\cdots,s_n) \tag{15-60}$$

它的含义是:在局势 $s=(s_1,\cdots,s_n)$中,局中人 i 将自己的策略由 s_i 换成 s_i^0,其他局中人的策略不变而得到的一个新局势。如果存在一个局势 s,使得对任意 $s_i^0\in s_i$,有

$$H_i(s)\geqslant H_i(s\parallel s_i^0)$$

则称局势 s 对局中人 i 有利,也就是说,若局势 s 对局中人 i 有利,则不论局中人 i 将自己的策略如何置换,都不会得到比在局势 s 下更多的赢得。显然,在非合作的条件下,每个局中人都力图选择对自己最有利的局势。

定义 8 如果局势 s 对所有的局中人都有利，即对任意 $i\in I, s_i^0\in S_i$，有

$$H_i(s)\geqslant H_i(s\parallel s_i^0) \tag{15-61}$$

则称 s 为非合作对策 G 的一个平衡局势(或平衡点)。

当 G 为二人零和对策时，上述定义等价为：$(\alpha_{i^*},\beta_{j^*})$为平衡局势的充要条件是：对任意 i,j，有

$$a_{ij^*}\leqslant a_{i^*j^*}\leqslant a_{i^*j} \tag{15-62}$$

此与前关于矩阵对策平衡局势的定义是一致的。

由矩阵对策的结果可知，非合作 n 人对策在纯策略意义下的平衡局势不一定存在。因此，需要考虑局中人的混合策略。对每个局中人的策略集 S_i，令 S_i^* 为定义在 S_i 上的混合策略集(即 S_i 上所有概率分布的集合)，x^i 表示局中人 i 的一个混合策略，$x=(x^1,\cdots,x^n)$为一个混合局势

$$x\parallel z^i=(x^1,\cdots,x^{i-1},z^i,x^{i+1},\cdots,x^n)$$

表示局中人 i 在局势 x 下，将自己的策略由 x^i 置换成 z^i 而得到的一个新的混合局势。以下，记 $E_i(x)$为局中人 i 在混合局势 x 下的赢得期望值，则有以下关于非合作 n 人对策解的定义。

定义 9 若对任意 $i\in I, z^i\in S_i^*$，有

$$E_i(x\parallel z^i)\leqslant E_i(x)$$

则称 x 为非合作 n 人对策 G 的一个平衡局势(或平衡点)。

对非合作 n 人对策，已经得到了一个非常重要的结论——定理 15。

定理 15 (纳什(Nash)定理) 非合作 n 人对策在混合策略意义下的平衡局势一定存在。

具体到二人有限非零和对策(亦称为双矩阵对策)，纳什定理的结论可表述为：一定存在 $x^*\in S_1^*, y^*\in S_2^*$，使得

$$x^{*\mathrm{T}}\boldsymbol{A}y^*\geqslant x^{\mathrm{T}}\boldsymbol{A}y^* \quad x\in S_1^* \tag{15-63}$$

$$x^{*\mathrm{T}}\boldsymbol{B}y^*\geqslant x^{*\mathrm{T}}\boldsymbol{B}y^* \quad y\in S_2^* \tag{15-64}$$

和矩阵对策所不同的是，双矩阵对策以及一般非合作 n 人对策平衡点的计算问题还远没有解决。但对 2×2 双矩阵对策，可得到如下结果：设双矩阵对策中两局中人的赢得矩阵分别为

$$\boldsymbol{A}=\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}\quad \boldsymbol{B}=\begin{bmatrix}b_{11} & b_{12}\\ b_{21} & b_{22}\end{bmatrix}$$

分别记局中人Ⅰ和Ⅱ的混合策略为$(x,1-x)$和$(y,1-y)$，由式(15-63)和式(15-64)，局势(x,y)的对策平衡点的充要条件是

$$E_1(x,y)\geqslant E_1(1,y) \tag{15-65}$$

$$E_1(x,y)\geqslant E_1(0,y) \tag{15-66}$$

$$E_2(x,y)\geqslant E_2(x,1) \tag{15-67}$$

$$E_2(x,y) \geqslant E_2(x,0) \tag{15-68}$$

由式(15-65),式(15-66),有

$$Q(1-x)y-q(1-x) \leqslant 0 \tag{15-69}$$

$$Qxy-qx \geqslant 0 \tag{15-70}$$

其中 $Q=a_{11}+a_{22}-a_{21}-a_{12}$,$q=a_{22}-a_{12}$,对式(15-69),式(15-70)求解,得到

(1) 当 $Q=0,q=0$ 时, $0\leqslant x\leqslant 1 \quad 0\leqslant y\leqslant 1$

(2) 当 $Q=0,q>0$ 时, $x=0 \quad 0\leqslant y\leqslant 1$

(3) 当 $Q=0,q<0$ 时, $x=1 \quad 0\leqslant y\leqslant 1$

(4) 当 $Q\neq 0$ 时,记 $q/Q=\alpha$,有

$$\begin{array}{ll} x=0 & y\leqslant\alpha \\ 0<x<1 & y=\alpha \\ x=1 & y\geqslant\alpha \end{array}$$

类似地,由式(15-67)和式(15-68),有

$$Rx(1-y)-r(1-y) \leqslant 0 \tag{15-71}$$

$$Rxy-ry \geqslant 0 \tag{15-72}$$

其中 $R=b_{11}+b_{22}-b_{21}-b_{12}$,$r=b_{22}-b_{21}$,对式(15-71),式(15-72)求解,得到

(1) 当 $R=0,r=0$ 时, $0\leqslant x\leqslant 1 \quad 0\leqslant y\leqslant 1$

(2) 当 $R=0,r>0$ 时, $0\leqslant x\leqslant 1 \quad y=0$

(3) 当 $R=0,r<0$ 时, $0\leqslant x\leqslant 1 \quad y=1$

(4) 当 $R\neq 0$ 时,记 $r/R=\beta$,有

$$\begin{array}{ll} x\leqslant\beta & y=0 \\ x=\beta & 0<y<1 \\ x\geqslant\beta & y=1 \end{array}$$

例 21 求解 2×2 双矩阵对策,其中

$$\boldsymbol{A}=\begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \quad \boldsymbol{B}=\begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}$$

解 由上面关于 2×2 双矩阵对策解的讨论,可知

$$Q=5>0 \quad q=2 \quad \alpha=q/Q=2/5$$

$$R=5>0 \quad r=3 \quad \beta=r/R=3/5$$

将这些结果代入双矩阵对策解的公式,得到

$$\begin{cases} x=0 & y\leqslant 2/5 \\ 0<x<1 & y=2/5 \\ x=1 & y\geqslant 2/5 \end{cases} \tag{15-73}$$

$$\begin{cases} x\leqslant 3/5 & y=0 \\ x=3/5 & 0<y<1 \\ x\geqslant 3/5 & y=1 \end{cases} \tag{15-74}$$

解不等式组(15-73)和不等式组(15-74)，得到对策的3个平衡点。

$$(x,y)=(0,0),\left(\frac{3}{5},\frac{2}{5}\right),(1,1)$$

不等式组(15-73)的解在图15-5中以粗线表示，不等式组(15-74)的解以虚线表示，粗线与虚线的3个交点即为对策的3个平衡点。

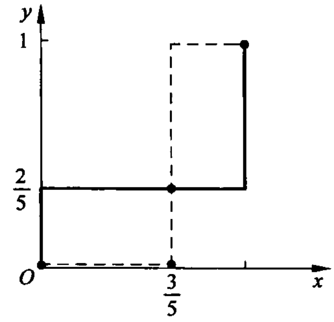


图 15-5

由

$$E_1(x,y)=5xy-2(x+y)+1 \tag{15-75}$$

$$E_2(x,y)=5xy-3(x+y)+2 \tag{15-76}$$

可得

$$E_1\left(\frac{3}{5},\frac{2}{5}\right)=E_2\left(\frac{3}{5},\frac{2}{5}\right)=\frac{1}{5} \tag{15-77}$$

$$E_1(0,0)=1,\quad E_1(1,1)=2 \tag{15-78}$$

$$E_2(0,0)=2,\quad E_2(1,1)=1 \tag{15-79}$$

不难发现，在平衡点(0,0)和(1,1)处，两个局中人的期望收益都比在平衡点(3/5,2/5)的期望收益要好。但由于这是一个非合作对策，不允许在选择策略前进行协商，所以两个局中人没有办法保证一定能达到平衡局势(0,0)或(1,1)。因而，尽管这个对策有3个平衡点，但哪一个平衡点作为对策的解都是难以令人信服的。

对例题“囚犯难题”(见本章例5)，不难得到两个囚犯的赢得矩阵分别为

$$\boldsymbol{A}=\begin{bmatrix}-7 & 0\\ -9 & -1\end{bmatrix},\boldsymbol{B}=\begin{bmatrix}-7 & -9\\ 0 & -1\end{bmatrix} \tag{15-80}$$

由

$$Q=1>0\quad q=-1\quad \alpha=q/Q=-1$$

$$R=1>0\quad r=-1\quad \beta=r/R=-1$$

不难确定该对策问题有唯一平衡点$(x,y)=(1,1)$，即两个人都承认犯罪，所得支付为各判刑7年。从赢得矩阵式(15-80)来看，这个平衡局势显然不是最有利的。如果两人都不承认犯罪，得到的赢得都是-1，相当于各判刑1年，这才是最有利的结果。但是，在非合作的条件下，这个最有利的结局也是难以达到的。

15.4.3 合作对策

由于非合作对策模型在适用性和理论上存在的局限性，使人们开始研究合作对策问题。合作对策的基本特征是参加对策的局中人可以进行充分的合作，即可以事先商定好，把各自的策略协调起来；可以在对策后对所得到的支付进行重新分配。合作的形式是所有局中人可以形成若干联盟，每个局中人仅参加一个联盟，联盟的所得要在联盟的所有成员中进行重新分配。一般说来，合作可以提高联盟的所得，因而也可以提高每个联盟成员

的所得。但联盟能否形成以及形成哪种联盟，或者说一个局中人是否参加联盟以及参加哪个联盟，不仅取决于对策的规则，更取决于联盟获得的所得如何在成员间进行合理的重新分配。如果分配方案不合理，就可能破坏联盟的形成，以至于不能形成有效的联盟。因此，在合作对策中，每个局中人如何选择自己的策略已经不是要研究的主要问题了，应当强调的是如何形成联盟，以及联盟的所得如何被合理地分配(即如何维持联盟)。

合作对策研究问题重点的转变，使得合作对策的模型、解的概念都和非合作对策问题有很大的不同。具体来说，构成合作对策的两个基本要素是：局中人集合 I 和特征函数 $v(S)$，其中 $I=\{1,2,\cdots,n\}$，S 为 I 的任一子集，也就是任何一个可能形成的联盟，$v(S)$ 表示联盟 S 在对策中的所得。合作对策的可行解是一个满足下列条件的 n 维向量 $x=(x_1,x_2,\cdots,x_n)$：

$$x_i \geqslant v(\{i\}) \quad i=1,\cdots,n \tag{15-81}$$

$$\sum_{i=1}^{n} x_i = v(I) \tag{15-82}$$

将满足式(15-81)和式(15-82)的向量 x 称为一个分配。合作对策研究的核心问题就是：如何定义“最优的”分配？是否存在“最优的”分配？怎样去求解“最优的”分配？鉴于对合作对策的系统阐述需要较多的数学知识和学时，超出了本书的基本要求，故不再作详细介绍，下面通过一个例子来说明合作对策的意义。

例 22 (产品定价问题) 设有两家厂商(厂商 1，厂商 2)为同一市场生产同样的产品，可选择的竞争策略是价格，目的是赚得最多的利润。已知两个厂商的需求函数为

$$Q_1=12-2P_1+P_2 \tag{15-83}$$

$$Q_2=12-2P_2+P_1 \tag{15-84}$$

其中 P_1，P_2 分别为两个厂商的价格，Q_1，Q_2 分别为市场对两个厂商产品的需求量(实际销售量)；又知，两家厂商的固定成本均为 20 元。于是，厂商 1 的利润函数为

$$\pi_1=P_1Q_1-20=12P_1-2P_1^2+P_1P_2-20 \tag{15-85}$$

为求厂商 1 利润最大化时的价格，令

$$\frac{\mathrm{d}\pi_1}{\mathrm{d}P_1}=12-4P_1+P_2=0 \tag{15-86}$$

得到

$$P_1=3+\frac{1}{4}P_2 \tag{15-87}$$

式(15-87)称为厂商 1 对厂商 2 的价格的反应函数；同理，可得到厂商 2 对厂商 1 的价格的反应函数为

$$P_2=3+\frac{1}{4}P_1 \tag{15-88}$$

由图 15-6 可以看出，如果两个厂商互不合作，各自从自身利润最大化出发，最稳妥的策略显然是都选择“定价 4 元”，也就是实现纳什均衡，各自可以得到 12 元的利润。但我们发现，如果两个厂商合作起来，都选择“定价 6 元”，则双方都可以赚得 16 元的利润，显然比

不合作时要好。因此,两个厂商可以结成一个"价格联盟",统一把价格定在 6 元,形成一个合作均衡,争取一个双赢的结果。

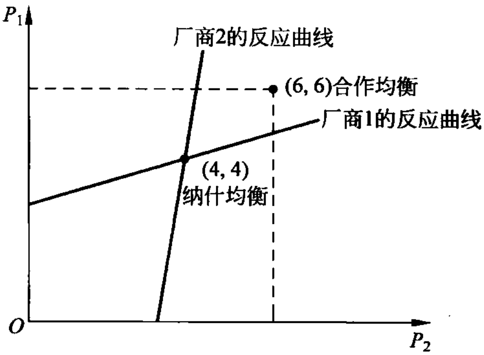


图 15-6 产品定价问题中的纳什均衡和合作均衡

但如果厂商 1 遵守价格联盟达成的合作协议,把价格定在 6 元,而厂商 2 却违反合作协议,将价格定在 4 元(即厂商 1 合作,而厂商 2 不合作),则厂商 1 的利润只有 4 元,而厂商 2 的利润却可以达到 20 元,如图 15-7 所示。

		厂商1定价 4元	厂商1定价 6元
厂商2定价	4元	12, 12	20, 4
厂商2定价	6元	4, 20	16, 16

图 15-7

这就给两个厂商带来了一个定价难题:到底采取哪个价格?一方面,"合作"的前景很诱人,但各厂商都担心,如果竞争对手不合作怎么办?而现实当中,一些厂商的确存在为了自身利益而违背市场竞争规则、与竞争对手进行削价竞争的冲动。不难看出,定价问题实际上正是"囚犯难题"在微观经济学中的一个实例。目前,"囚犯难题"的模型已被应用于经济学、社会学、心理学、伦理学、政治学等众多领域,充分说明由此模型而引出的合作对策模型具有十分广泛的适应性和应用前景。

习题

15.1 甲、乙两个儿童玩游戏,双方可分别出拳头(代表石头)、手掌(代表布)、两个手指(代表剪刀),规则是:剪刀赢布,布赢石头,石头赢剪刀,赢者得一分。若双方所出相同算和局,均不得分。试列出儿童甲的赢得矩阵。

15.2 "二指莫拉问题"。甲、乙二人游戏,每人出一个或两个手指,同时又把猜测对方所出的指数叫出来。如果只有一个人猜测正确,则他所赢得的数目为二人所出指数之和,否则重新开始,写出该对策中各局中人的策略集合及甲的赢得矩阵,并回答局中人是否存在某种出法比其他出法更为有利。

15.3 求解下列矩阵对策,其中赢得矩阵 A 分别为

(1) $\begin{bmatrix} -2 & 12 & -4 \\ 1 & 4 & 8 \\ -5 & 2 & 3 \end{bmatrix}$　　(2) $\begin{bmatrix} 2 & 2 & 1 \\ 3 & 4 & 4 \\ 2 & 1 & 6 \end{bmatrix}$

(3) $\begin{bmatrix} 2 & 7 & 2 & 1 \\ 2 & 2 & 3 & 4 \\ 3 & 5 & 4 & 4 \\ 2 & 3 & 1 & 6 \end{bmatrix}$　　(4) $\begin{bmatrix} 9 & 3 & 1 & 8 & 0 \\ 6 & 5 & 4 & 6 & 7 \\ 2 & 4 & 3 & 3 & 8 \\ 5 & 6 & 2 & 2 & 1 \\ 3 & 2 & 3 & 5 & 4 \end{bmatrix}$

15.4 证明本章 15.2.1 中的性质 1 和性质 2。

15.5 甲、乙两个企业生产同一种电子产品,两个企业都想通过改革管理获取更多的市场销售份额。

甲企业的策略措施有:① 降低产品价格;② 提高产品质量,延长保修年限;③ 推出新产品。

乙企业考虑的措施有:① 增加广告费用;② 增设维修网点,扩大维修服务;③ 改进产品性能。

假定市场份额一定,由于各自采取的策略措施不同,通过预测,今后两个企业的市场占有份额变动情况如表 15-6 所示(正值为甲企业增加的市场占有份额,负值为减少的市场占有份额)。试通过对策分析,确定两个企业各自的最优策略。

表 15-6

甲企业策略	乙企业策略		
	1	2	3
1	10	−1	3
2	12	10	−5
3	6	8	5

15.6 证明本章中的定理 2。

15.7 证明本章中的定理 4。

15.8 证明本章中的定理 7、定理 8 和定理 9。

15.9 设 G 为 2×2 对策,且不存在鞍点。证明若 $x^*=(x_1^*,x_2^*)^{\mathrm{T}}$ 和 $y^*=(y_1^*,y_2^*)^{\mathrm{T}}$ 是 G 的解,则

$$x_i^*>0 \quad i=1,2$$
$$y_j^*>0 \quad j=1,2$$

15.10* 证明:矩阵对策

$$\boldsymbol{A}=\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

的鞍点不存在的充要条件是有一条对角线的每一个元素均大于另一对角线上的每一元素。

15.11* 推导 2×2 对策的求解公式(15-25)~式(15-29)，并由习题 15.10 的结果证明习题 15.9。

15.12 设 $m\times m$ 对策的矩阵为

$$\boldsymbol{A}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix}$$

其中当 $i\neq j$ 时，$a_{ij}=1$，当 $i=j$ 时，$a_{ii}=-1$，证明此对策的最优策略为

$$x^*=y^*=(1/m,1/m,\cdots,1/m)^{\mathrm{T}}$$

$$V_G=\frac{m-2}{m}$$

15.13 利用优超原则求解下列矩阵对策

(1) $\boldsymbol{A}=\begin{bmatrix} 1 & 0 & 3 & 4 \\ -1 & 4 & 0 & 1 \\ 2 & 2 & 2 & 3 \\ 0 & 4 & 1 & 1 \end{bmatrix}$　　(2) $\boldsymbol{A}=\begin{bmatrix} 3 & 4 & 0 & 3 & 0 \\ 5 & 0 & 2 & 5 & 9 \\ 7 & 3 & 9 & 5 & 9 \\ 4 & 6 & 8 & 7 & 6 \\ 6 & 0 & 8 & 8 & 3 \end{bmatrix}$

15.14 利用图解法求解下列矩阵对策，其中矩阵 $\boldsymbol{A}$ 分别为

(1) $\begin{bmatrix} 2 & 4 \\ 2 & 3 \\ 3 & 2 \\ -2 & 6 \end{bmatrix}$　　(2) $\begin{bmatrix} 5 & 7 & -6 \\ -6 & 0 & 4 \\ 7 & 8 & 5 \end{bmatrix}$

(3) $\begin{bmatrix} 4 & 2 & 3 & -1 \\ -4 & 0 & -2 & -2 \end{bmatrix}$　　(4) $\begin{bmatrix} 1 & 3 & 11 \\ 8 & 5 & 2 \end{bmatrix}$

15.15 已知矩阵对策

$$\boldsymbol{A}=\begin{bmatrix} 4 & 0 & 0 \\ 0 & 0 & 8 \\ 0 & 6 & 0 \end{bmatrix}$$

的解为 $x^*=(6/13,3/13,4/13)^{\mathrm{T}}$，$y^*=(6/13,4/13,3/13)^{\mathrm{T}}$，对策值为 24/13。求下列矩阵对策的解，其赢得矩阵 $\boldsymbol{A}$ 分别为

(1) $\begin{bmatrix} 6 & 2 & 2 \\ 2 & 2 & 10 \\ 2 & 8 & 2 \end{bmatrix}$　　(2) $\begin{bmatrix} -2 & -2 & 2 \\ 6 & -2 & -2 \\ -2 & 4 & -2 \end{bmatrix}$　　(3) $\begin{bmatrix} 32 & 20 & 20 \\ 20 & 20 & 44 \\ 20 & 38 & 20 \end{bmatrix}$

15.16　**用线性规划方法求解下列矩阵对策，其中 $\boldsymbol{A}$ 为**

(1) $\begin{bmatrix} 8 & 2 & 4 \\ 2 & 6 & 6 \\ 6 & 4 & 4 \end{bmatrix}$　　(2) $\begin{bmatrix} 2 & 0 & 2 \\ 0 & 3 & 1 \\ 1 & 2 & 1 \end{bmatrix}$

参考资料

[1]　中国科学院数学研究所第二室.对策论(博弈论)讲义.北京：人民教育出版社,1960

[2]　J.麦克金赛,高鸿勋,曾鼎钚,王厦生译.博弈论导引.北京：人民教育出版社,1960

[3]　张盛开.对策论及其应用.武汉：华中工学院出版社,1985

[4]　Owen G. Game Theory. Academic Press,1982

Part 10

第 10 篇

决策论

决策是人们在政治、经济、技术以及日常生活中普遍遇到的一种选择方案的行为。其困难是如何从多种方案中作出正确的选择，以便获得好的结果或达到预期的目标。管理国家、企业时刻都遇到大大小小的决策问题。诺贝尔奖金获得者西蒙(A. Simon)有一句名言："管理就是决策。"就是说管理的核心是决策。决策的失误是最大的失误。将会造成巨大的损失。人们必须冥思苦想地学会正确的决策。决策是由直观经验决策发展到科学民主的决策。将现代科学技术和现代文明的成就应用于研究决策，称为决策科学。决策科学是在研究决策活动基本规律的基础上，总结出一套进行决策必须遵循的原则、规则、程序、方法和技术。

研究决策的问题包括：决策的基本原理、决策的程序、决策的信息、决策的方法(定量与定性的方法)、决策的风险、决策中的人因素、决策的思维方式、决策的组织、决策的实施等。决策科学包括的内容十分广泛：涉及社会学、决策心理学、决策行为学、决策的量化方法和评价、决策支持系统和决策自动化等多学科和多领域的综合应用。

近十年来，决策科学的研究发展很快，在单人单目标决策、单人多目标决策方面，特别是群决策方面(群决策或群体决策是指参与决策的人数多于一人的决策行为)，引起理论工作者与实践工作者的广泛关注。

这里仅从运筹学中的定量分析方法的角度向读者作入门的介绍。

CHAPTER 16
第16章 单目标决策

16.1 决策的分类

从不同的角度出发可得不同的决策分类。

1. 按性质的重要性分类

可将决策分为战略决策、策略决策和执行决策，或叫战略计划、管理控制和运行控制。

战略决策是涉及某组织发展和生存有关的全局性、长远问题的决策，如厂址的选择、新产品开发方向、新市场的开发、原料供应地的选择等。

策略决策是为完成战略决策所规定的目的而进行的决策，如对一个企业产品规格的选择、工艺方案和设备的选择、厂区和车间内工艺路线的布置等。

执行决策是根据策略决策的要求对执行行为方案的选择，如生产中产品合格标准的选择、日常生产调度的决策等。

2. 按决策的结构分类

分为程序决策和非程序决策。

程序决策是一种有章可循的决策，一般是可重复的。

非程序决策一般是无章可循的决策，只能凭经验直觉做出应变的决策，一般是一次性的。

由于决策的结构不同，解决问题的方式也不同，现归纳于表16-1。

表 16-1

解决问题的方式	程序决策	非程序决策
传统方式	习惯，标准规程	直观判断，创造性观察，选拔人才
现代方式	运筹学，管理信息系统	培训决策者，人工智能，专家系统

3. 按定量和定性分类

分为定量决策和定性决策，描述决策对象的指标都可以量化时可用定量决策，否则只能用定性决策。总的发展趋势是尽可能地把决策问题量化。

4. 按决策环境分类

可将决策问题分为确定型的、风险型的和不确定型的三种。确定型的决策是指决策

环境是完全确定的,作出的选择的结果也是确定的。风险型决策是指决策的环境不是完全确定的,而其发生的概率是已知的。不确定型决策是指决策者对将发生结果的概率一无所知,只能凭决策者的主观倾向进行决策。

5. 按决策过程的连续性分类

可分为单项决策和序贯决策。单项决策是指整个决策过程只作一次决策就得到结果,序贯决策是指整个决策过程由一系列决策组成,一般讲管理活动是由一系列决策组成的,但在这一系列决策中往往有几个关键环节要作决策,可以把这些关键的决策分别看做单项决策。

16.2 决策过程

构造人们决策行为的模型主要有两种方法:一种是面向决策结果的方法;另一种是面向决策过程的方法。

面向决策结果的方法认为:若决策者能正确地预见到决策结果,其核心是决策的结果和正确的预测。通常的单目标决策和多目标决策是属于这种类型的。

面向决策过程的方法认为:若决策者了解了决策过程,掌握了过程和能控制过程,他就能正确地预见决策的结果。对于面向决策结果的方法的程序比较简单,见图 16-1。



图 16-1

由图 16-1 可知,任何决策都有一个过程和程序,绝非决策者灵机一动拍板就行。面向决策过程的方法一般包括预决策→决策→决策后三个互相依赖的阶段。

预决策阶段是指当要决策的问题摆在决策者面前时,决策者能立即想到各种可能方案,并意识到没有理想方案时,就会产生矛盾。他开始企图寻找减少矛盾的方案,沿着这企图扩大线索时,就需要收集信息。收集信息时开始时比较客观,无倾向性,以后逐渐变得主观和有倾向了。当预决策进行得较顺利,可以进行局部决策。

决策阶段可分为分部决策和最终决策两个阶段。分部决策包括对决策处境作方向性的调整,如排除劣解,重新考虑已放弃的方案,增加和去掉一些评价准则。在合并一些方案后,减少了变量数和方案数,决策者按主观倾向重新估价各方案,并保留倾向的少数方案,以便进行最终决策。

决策后阶段,当进行了最终决策,这时主要考虑的问题是决策后看法不一致。这时决策者倾向于解释和强调已选方案的优点,并寻找更多的信息来证明已选方案的优点和正确性。应克服愿听取相同意见、不愿听取不同意见的现象。决策后阶段要对决策实施进行了解,这是十分重要的,决策实施是决策的继续,决策后阶段往往也是下次决策的预决

策阶段。

任何决策问题都由以下要素构成决策模型

(1) 决策者,他的任务是进行决策。决策者可以是个人、委员会或某个组织。一般指领导者或领导集体。

(2) 可供选择的方案(替代方案)、行动或策略。参谋人员的任务是为决策者提供各种可行方案。这里包括了解研究对象的属性,确定目的和目标。

属性是指研究对象的特性,它们是客观存在的,是可以客观量度的,并由决策者主观选定的,如选拔飞行员时,按身高、年龄、健康状况等数值来表明其属性。

目的是表明选择属性的方向,如要优秀还是良好,反映了决策者的要求和愿望。

目标是给出了参数值的目的,如目的是选择一种省油的汽车时,那么以每公升能行驶 60 千米为目标。

(3) 准则是衡量选择方案,包括目的、目标、属性、正确性的标准,在决策时有单一准则和多准则。

(4) 事件是指不为决策者所控制的客观存在的将发生的状态。

(5) 每一事件的发生将会产生某种结果,如获得收益或损失。

(6) 决策者的价值观,如决策者对货币额或不同风险程度的主观价值观念。

确定型的决策是指不包含随机因素的决策问题,每个决策都会得到一个唯一的事先可知的结果。从决策论的观点来看,前几章讨论的规划论等都是确定型的决策问题。本章讨论的决策问题都是具有不确定因素和有风险的决策。

16.3 不确定型的决策

所谓不确定型的决策是指决策者对环境情况一无所知。这时决策者是根据自己的主观倾向进行决策,由决策者的主观态度不同基本可分为四种准则:悲观主义决策准则、乐观主义决策准则、等可能性准则、最小机会损失决策准则。以下用例子分别说明之。

例 1 设某工厂是按批生产某产品并按批销售,每件产品的成本为 30 元,批发价格为每件 35 元。若每月生产的产品当月销售不完,则每件损失 1 元。工厂每投产一批是 10 件,最大月生产能力是 40 件,决策者可选择的生产方案为 0 件、10 件、20 件、30 件、40 件五种。假设决策者对其产品的需求情况一无所知,试问这时决策者应如何决策?

这个问题可用决策矩阵来描述。决策者可选的行动方案有五种,这是他的策略集合,记作 $\{S_i\}$, $i=1,2,\cdots,5$。经分析他可断定将发生五种销售情况:即销量为 0 件,10 件,20 件,30 件,40 件,但不知它们发生的概率。这就是事件集合,记作 $\{E_j\}$, $j=1,2,\cdots,5$。每个"策略—事件"对都可以计算出相应的收益值或损失值。如当选择月产量为 20 件时,而销出量为 10 件,这时收益额为

$$10\times(35-30)-1\times(20-10)=40(\text{元})$$

可以一一计算出各“策略—事件”对应的收益值或损失值，记作 a_{ij}。将这些数据汇总在矩阵中，见表 16-2。

表 16-2

S_i		事件 E_j				
		0	10	20	30	40
策略	0	0	0	0	0	0
	10	−10	50	50	50	50
	20	−20	40	100	100	100
	30	−30	30	90	150	150
	40	−40	20	80	140	200

这就是决策矩阵。根据决策矩阵中元素所示的含义不同，可称为收益矩阵、损失矩阵、风险矩阵、后悔值矩阵等。

下面讨论决策者是如何应用决策准则进行决策的。

16.3.1 悲观主义(max min)决策准则

悲观主义决策准则亦称保守主义决策准则。当决策者面临着各事件的发生概率不清时，决策者考虑可能由于决策错误而造成重大经济损失。由于自己的经济实力比较脆弱，他在处理问题时就较谨慎。他分析各种最坏的可能结果，从中选择最好者，以它对应的策略为决策策略，用符号表示为 max min 决策准则。在收益矩阵中先从各策略所对应的可能发生的“策略—事件”对的结果中选出最小值，将它们列于表的最右列。再从此列的数值中选出最大者，以它对应的策略为决策者应选的决策策略。计算见表 16-3。

表 16-3

S_i		事件 E_j					min
		0	10	20	30	40	
策略	0	0	0	0	0	0	0←max
	10	−10	50	50	50	50	−10
	20	−20	40	100	100	100	−20
	30	−30	30	90	150	150	−30
	40	−40	20	80	140	200	−40

根据 max min 决策准则有

$$\max\ (0,-10,-20,-30,-40)=0$$

它对应的策略为 S_1，即为决策者应选的策略。在这里是“什么也不生产”，这结论似乎荒谬，但在实际中表示先看一看，以后再作决定。上述计算用公式表示为

$$S_k^* \rightarrow \max\ \min\ (a_{ij})$$

16.3.2 乐观主义(max max)决策准则

持乐观主义(max max)决策准则的决策者对待风险的态度与悲观主义者不同,当他面临情况不明的策略问题时,他绝不放弃任何一个可获得最好结果的机会,以争取好中之好的乐观态度来选择他的决策策略。决策者在分析收益矩阵各策略的"策略—事件"对的结果中选出最大者,记在表的最右列。再从该列数值中选择最大者,以它对应的策略为决策策略,见表 16-4。

表 16-4

S_i		事件 E_j					max
		0	10	20	30	40	
策略	0	0	0	0	0	0	0
	10	−10	50	50	50	50	50
	20	−20	40	100	100	100	100
	30	−30	30	90	150	150	150
	40	−40	20	80	140	200	200←max

根据 max max 决策准则有

$$\max\ (0,50,100,150,200)=200$$

它对应的策略为 S_5。用公式表示为

$$S_k^* \rightarrow \max_i \max_j (a_{ij})$$

16.3.3 等可能性(Laplace)准则

等可能性(Laplace)准则是 19 世纪数学家 Laplace 提出的。他认为:当一个人面临着某事件集合,在没有什么确切理由来说明这一事件比那一事件有更多发生机会时,只能认为各事件发生的机会是均等的。即每一事件发生的概率都是 1/事件数。决策者计算各策略的收益期望值,然后在所有这些期望值中选择最大者,以它对应的策略为决策策略,见表 16-5。然后按下式决定决策策略。

$$S_k^* \rightarrow \max_i \{E(S_i)\}$$

表 16-5

S_i		事件 E_j					$E(S_i)=\sum_j pa_{ij}$
		0	10	20	30	40	
策略	0	0	0	0	0	0	0
	10	−10	50	50	50	50	38
	20	−20	40	100	100	100	64
	30	−30	30	90	150	150	78
	40	−40	20	80	140	200	80←max

在本例中 $P=\frac{1}{5}$,期望值

$$E(S_i)=\sum_j pa_{ij}$$

$$\max\{E(S_i)\}=\max\{0,38,64,78,80\}=80$$

它对应的策略 S_5 为决策策略。

16.3.4 最小机会损失决策准则

最小机会损失决策准则亦称最小遗憾值决策准则或 Savage 决策准则。首先将收益矩阵中各元素变换为每一"策略—事件"对的机会损失值(遗憾值,后悔值)。其含义是:当某一事件发生后,由于决策者没有选用收益最大的策略,而形成的损失值。若发生 k 事件,各策略的收益为 a_{ik},$i=1,2,\cdots,5$,其中最大者为

$$a_{ik}=\max_i(a_{ik})$$

这时各策略的机会损失值为

$$a'_{ik}=\{\max_i(a_{ik})-a_{ik}\},i=1,\cdots,5$$

计算结果见表 16-6。

表 16-6

S_i		事件 E_j 0	10	20	30	40	max
策略	0	0	50	100	150	200	200
	10	10	0	50	100	150	150
	20	20	10	0	50	100	100
	30	30	20	10	0	50	50
	40	40	30	20	10	0	40←min

从所有最大机会损失值中选取最小者,它对应的策略为决策策略。用公式表示为

$$S_k^* \to \min_i \max_j a'_{ij}$$

本例的决策策略为

$$\min(200,150,100,50,40)=40\to S_5$$

在分析产品废品率时,应用本决策准则就比较方便。

16.3.5 折中主义准则

当用 min max 决策准则或 max max 决策准则来处理问题时,有的决策者认为这样太极端了。于是提出把这两种决策准则给予综合,令 a 为乐观系数,且 $0\leqslant a\leqslant 1$。并用以下关系式表示

$$H_i = \alpha a_{i_{max}} + (1-\alpha) a_{i_{min}}$$

$a_{i_{max}}$, $a_{i_{min}}$ 分别表示第 i 个策略可能得到的最大收益值与最小收益值。设 $\alpha=1/3$，将计算得的 H_i 值记在表 16-7 的右端。

然后选择

$$S_k^* \to \max_i \{H_i\}$$

本例的决策策略为

$$\max\ (0,10,20,30,40)=40 \to S_5$$

表 16-7

S_i		事件 E_j					H_i
		0	10	20	30	40	
	0	0	0	0	0	0	0
策	10	−10	50	50	50	50	10
	20	−20	40	100	100	100	20
略	30	−30	30	90	150	150	30
	40	−40	20	80	140	200	40←max

在不确定性决策中是因人、因地、因时选择决策准则的，但在实际中当决策者面临不确定性决策问题时，他首先是获取有关各事件发生的信息，使不确定性决策问题转化为风险决策，风险决策将是讨论的重点。

16.4 风险决策

风险决策是指决策者对客观情况不甚了解，但对将发生各事件的概率是已知的。决策者往往通过调查，根据过去的经验或主观估计等途径获得这些概率。在风险决策中一般采用期望值作为决策准则，常用的有最大期望收益决策准则和最小机会损失决策准则。

16.4.1 最大期望收益决策准则(expected monetary value, EMV)

决策矩阵的各元素代表“策略—事件”对的收益值，各事件发生的概率为 p_j。先计算各策略的期望收益值

$$\sum_j p_j a_{ij}, \qquad i=1,2,\cdots,n$$

然后从这些期望收益值中选取最大者，它对应的策略为决策应选策略。即

$$\max_i \sum_j p_j a_{ij} \to S_k^*$$

以例 1 的数据进行计算，见表 16-8。

表 16-8

S_i		事件 E_j 0	10	20	30	40	EMV
		p_j 0.1	0.2	0.4	0.2	0.1	
	0	0	0	0	0	0	0
策	10	−10	50	50	50	50	44
	20	−20	40	100	100	100	76
略	30	−30	30	90	150	150	84←max
	40	−40	20	80	140	200	80

这时

$$\max(0,44,76,84,80)=84 \rightarrow S_4$$

即选择策略 $S_4=30$。

EMV 决策准则适用于一次决策多次重复进行生产的情况，所以它是平均意义下的最大收益。

16.4.2 最小机会损失决策准则(expected opportunity loss，EOL)

矩阵的各元素代表“策略—事件”对的机会损失值，各事件发生的概率为 p_j，先计算各策略的期望损失值。

$$\sum_j p_j a'_{ij}, \qquad i=1,2,\cdots,n$$

然后从这些期望损失值中选取最小者，它对应的策略应是决策者所选策略。即

$$\min_i\left(\sum_j p_j a'_{ij}\right) \rightarrow S_k^*$$

表上运算与上述相似。

16.4.3 EMV 与 EOL 决策准则的关系

从本质上讲 EMV 与 EOL 决策准则是一样的。

设 a_{ij} 为决策矩阵的收益值。因为当发生的事件的所需量等于所选策略的生产量时，收益值最大，即在收益矩阵的对角线上的值都是其所在列中的最大者。于是机会损失矩阵可通过以下求得，见表 16-9。

表 16-9

S_i	E_1	E_2	…	E_n
	p_1	p_2	…	p_n
S_1	$a_{11}-a_{11}$	$a_{22}-a_{12}$	…	$a_{nn}-a_{1n}$
S_2	$a_{11}-a_{21}$	$a_{22}-a_{22}$	…	$a_{nn}-a_{2n}$
⋮	⋮	⋮	⋮	⋮
S_n	$a_{11}-a_{n1}$	$a_{22}-a_{n2}$	…	$a_{nn}-a_{nn}$

第 i 策略的机会损失

$$\begin{aligned} EOL_i &= p_1(a_{11}-a_{i1})+p_2(a_{22}-a_{i2})+\cdots+p_n(a_{nn}-a_{in}) \\ &= p_1a_{11}+p_2a_{22}+\cdots+p_na_{nn}-(p_1a_{i1}+p_2a_{i2}+\cdots+p_na_{in}) \\ &= K-(p_1a_{i1}+p_2a_{i2}+\cdots+p_na_{in}) \\ &= K-EMV_i \end{aligned}$$

其中，$K=p_1a_{11}+p_2a_{22}+\cdots+p_na_{m}$ 与 i 无关，而 $p_1a_{i1}+p_2a_{i2}+\cdots+p_na_{in}=EMV_i$。故当 EMV 为最大时，EOL 便为最小。所以在决策时用这两个决策准则所得结果是相同的。

16.4.4 全情报的价值(expected value of perfactinformation，EVPI)

当决策者耗费了一定经费进行调研，获得了各事件发生概率的信息，应采用“随机应变”的战术。这时所得的期望收益称为全情报的期望收益记作 EPPL。这收益应当大于至少等于最大期望收益，即 $EPPL \geqslant EMV^*$。则

$$EPPL-EMV^*=EVPI$$

称为对全情报的价值。这就是说明获取情报的费用不能超过 EVPI 值，否则就没有增加收入。

实际应用时考虑费用构成很复杂，这里仅说明全情报价值的概念和其意义。

16.4.5 主观概率

风险决策时决策者要估计各事件出现的概率，而许多决策问题的概率不能通过随机试验去确定，根本无法进行重复试验。如估计某企业倒闭的可能性，只能由决策者根据他对这事件的了解去确定。这样确定的概率反映了决策者对事件出现的信念程度，称为主观概率。客观概率论者认为概率如同重量、容积、硬度等一样，是研究对象的物理属性。而主观概率论者则认为概率是人们对现象的知识的现状的测度，而不是现象本身的测度，因此不是研究对象的物理属性。主观概率论者不是主观臆造事件发生的概率，而是依赖于对事件做周密的观察，去获得事前信息。事前信息愈丰富，则确定的主观概率就愈准确。主观概率论者并不否认实践是第一性的观点。所以主观概率是进行决策的依据。确定主观概率时，一般采用专家估计法。

1. 直接估计法

直接估计法是要求参加估计者直接给出概率的估计方法。

例如推荐三名大学生考研究生时，请五位任课教师估计他们谁得第一的概率。若各任课教师作出如下的估计，见表 16-10。由表 16-10 的末行得到学生 1 的概率是 0.47，他是最高者。

表 16-10

教师代号	权数	学生1	学生2	学生3	Σ
1	0.6	0.6	0.6	0.1	
2	0.7	0.4	0.5	0.1	
3	0.9	0.5	0.3	0.2	
4	0.7	0.6	0.3	0.1	
5	0.8	0.2	0.5	0.3	
		1.67	1.31	0.55	3.53
归一化后		0.47	0.37	0.16	1

2. 间接估计法

参加估计者通过排队或相互比较等间接途径给出概率的估计方法。

例如估计五个球队($A_i, i=1,\cdots,5$)比赛谁得第一的问题，请十名专家作出估计，每位都给出一个优胜顺序的排列名单，排队名单汇总在表 16-11。

表 16-11

专家号	名次 q_j					评定者
	1	2	3	4	5	权数 w_i
1	A_2	A_5	A_1	A_3	A_4	0.7
2	A_3	A_1	A_5	A_4	A_2	0.8
3	A_5	A_3	A_2	A_1	A_4	0.6
4	A_1	A_2	A_5	A_4	A_3	0.7
5	A_5	A_2	A_1	A_3	A_4	0.9
6	A_2	A_5	A_3	A_1	A_4	0.8
7	A_5	A_1	A_3	A_2	A_4	0.7
8	A_5	A_2	A_4	A_1	A_3	0.9
9	A_2	A_1	A_5	A_4	A_3	0.7
10	A_5	A_2	A_3	A_1	A_4	0.8

分别从表 16-11 查得每队被排的名次的次数，如 A_1 所处各名次的意见为

q_j	次数 n_j	评定权数 w_i
1	1	$w_4=0.7$
2	3	$w_2=0.8, w_7=0.7, w_9=0.7$
3	2	$w_1=0.7, w_5=0.9$
4	4	$w_3=0.6, w_6=0.8, w_8=0.9, w_{10}=0.8$
5	0	

然后计算加权平均数

$$w(A_1)=\frac{1\times w_4+2(w_2+w_7+w_9)+3(w_1+w_5)+4(w_3+w_6+w_8+w_{10})}{\sum w_i}$$

$$=22.3/7.6=2.93$$

采用同样方法得到

$$w(A_2)=(17.4/7.6)=2.29;\ w(A_3)=(26.8/7.6)=3.53;$$

$$w(A_4)=(34/7.6)=4.47;\ w(A_5)=(13.5/7.6)=1.78$$

这就可以按此加权平均数给出各队的估计名次，即

$$A_5>A_2>A_1>A_3>A_4$$

下面再将各队的估计名次转换成概率，这时需假设各队按估计名次出现的概率是等可能的。$(A_5\rightarrow 1)$表示 A_5 的估计名次为 1，其余类推。那么

$$(A_5\rightarrow 1):(A_2\rightarrow 2):(A_1\rightarrow 3):(A_3\rightarrow 4):(A_4\rightarrow 5)=1:1:1:1:1$$

因所有事件发生的概率和为 1，即

$$\sum_j p_j=1$$

于是各队按估计名次出现的主观概率为

$$P(A_5\rightarrow 1)=1/5;\quad P(A_2\rightarrow 2)=1/5;\quad P(A_1\rightarrow 3)=1/5$$

$$P(A_3\rightarrow 4)=1/5;\quad P(A_4\rightarrow 5)=1/5$$

当然决策者还可根据了解的情况，作其他的假设，这样就能得到另外的结果。

16.4.6 修正概率的方法——贝叶斯公式的应用

前面曾提到决策者常常碰到的问题是没有掌握充分的信息，于是决策者通过调查及做试验等途径去获得更多的更确切的信息，以便掌握各事件发生的概率，这可以利用贝叶斯公式来实现，它体现了最大限度的利用现有信息，并加以连续观察和重新估计，其步骤为：

① 先由过去的经验或专家估计获得将发生事件的事前(先验)概率。

② 根据调查或试验计算得到条件概率，利用贝叶斯公式：

$$P(B_i\mid A)=\frac{P(B_i)P(A\mid B_i)}{\sum P(B_i)P(A\mid B_i)}\qquad i=1,2,\cdots,n$$

计算出各事件的事后(后验)概率。

例 2 某钻探大队在某地区进行石油勘探，主观估计该地区有油的概率为$P(O)=0.5$；无油的概率为$P(D)=0.5$。为了提高钻探的效果，先做地震试验。根据积累的资料得知：凡有油地区做试验结果好的概率为 $P(F|O)=0.9$；做试验结果不好的概率为 $P(U|O)=0.1$。凡无油地区做试验结果好的概率为 $P(F|D)=0.2$；做试验结果不好的概率为 $P(U|D)=0.8$。问在该地区做试验后，有油与无油的概率各是多少？

解 先计算做地震试验好与不好的概率。

做地震试验好的概率

$$P(F)=P(O)\cdot P(F|O)+P(D)\cdot P(F|D)$$
$$=0.5\times0.9+0.5\times0.2=0.55$$

做地震试验不好的概率

$$P(U)=P(O)\cdot P(U|O)+P(D)\cdot P(U|D)$$
$$=0.5\times0.8+0.5\times0.1=0.45$$

利用贝叶斯公式计算各事件的事后(后验)概率。

做地震试验好的条件下有油的概率

$$P(O|F)=\frac{P(O)\cdot P(F|O)}{P(F)}=\frac{0.45}{0.55}=\frac{9}{11}$$

做地震试验好的条件下无油的概率

$$P(D|F)=\frac{P(D)\cdot P(F|D)}{P(F)}=\frac{0.10}{0.55}=\frac{2}{11}$$

做地震试验不好的条件下有油的概率

$$P(O|U)=\frac{P(O)\cdot P(U|O)}{P(U)}=\frac{0.05}{0.45}=\frac{1}{9}$$

做地震试验不好的条件下无油的概率

$$P(D|U)=\frac{P(D)\cdot P(U|D)}{P(U)}=\frac{0.40}{0.45}=\frac{8}{9}$$

以上计算可在图 16-2 上进行。

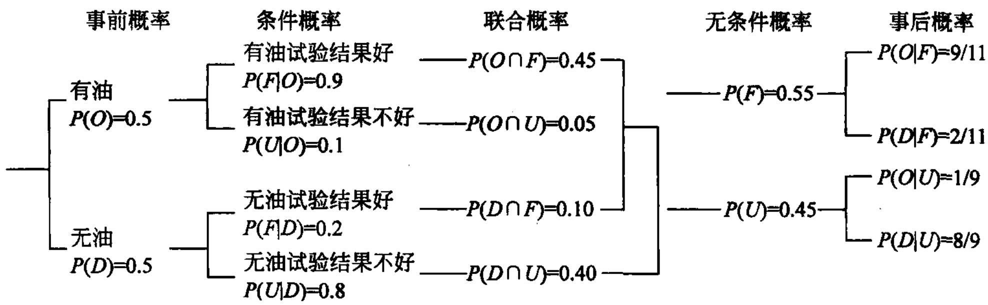


图 16-2

例 3 某厂生产电子元件,每批的次品率的概率分布见表 16-12。该厂不进行 100% 的检验,现抽样 20 件,次品为 1 件,试修订事前概率。

表 16-12

次品率 p	0.02	0.05	0.10	0.15	0.20
事前概率 $p_0(p)$	0.4	0.3	0.15	0.10	0.05

解 为了便于计算,将表 16-12 的数据填入表 16-13 的(1)、(2)列中。

表 16-13

(1) 次品率 p	(2) 事前概率 $P_0(p)$	(3) 条件概率 $P(x=1\|20,p)$	(4) 联合概率 $P(x=1\cap p)$	(5)事后概率 $P(p\|x=1)$
0.02	0.4	0.272 5	0.109 00	0.390 30
0.05	0.3	0.377 4	0.113 19	0.405 31
0.10	0.15	0.270 1	0.040 52	0.145 09
0.15	0.10	0.136 8	0.013 68	0.048 99
0.20	0.05	0.057 7	0.002 88	0.010 31
合　计	1.00		$P(x=1)=0.279\ 27$	1.000 0

(3)列的数字表示在次品率为 P 的母体中抽 20 个检验，有 1 个次品的概率，这概率可用以下计算得到。因产品抽样检验的次品率是服从二项分布的，可得到

$$P(x|n,p)=\frac{n!}{x!\ (n-x)!}p^x q^{n-x}$$

可用 MS. Excel 计算或查表得到

$$p(x=1|20,0.02)=0.272\ 5$$
$$p(x=1|20,0.05)=0.377\ 4$$
$$\cdots\cdots$$

(4)列的数字是按(4)=(2)×(3)求得的。然后求

$$P(x=1)=\sum P(x=1\cap p_i)=0.279\ 27$$

事后概率按$(5)=\dfrac{(4)}{0.279\ 27}$求得。

由以上两例可见，求修正概率时，可用树形图或表格计算。

16.5 效用理论在决策中的应用

16.5.1 效用及效用曲线

效用概念首先是由贝努里(D. Berneulli)提出的，他认为人们对其钱财的真实价值的考虑与他的钱财拥有量之间有对数关系。如图 16-3 所示，这就是贝努里的货币效用函数。经济管理学家将效用作为指标，用它来衡量人们对某些事物的主观价值、态度、偏爱、倾向等。例如在风险情况下进行决策，决策者对风险的态度是不同的。用效用指标来量化决策者对待风险的态度，可以给每个决策者测定他的对待风险的态度的效用曲线(函数)。效用值是一个相对的指标值，一般可规定：凡对决策者最爱好、最倾向、最愿意的事物(事件)的效用值赋予 1；而最不爱好的赋予 0。也可以用其他数值范围，如(100～0)。这

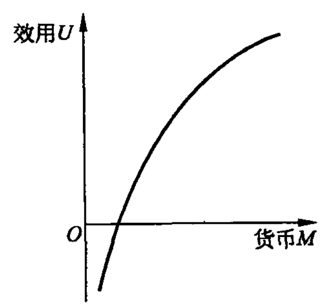


图 · 16-3

如同水的冰点可以用0℃表示或用32℉表示，但效用无量纲指标。通过效用这指标可将某些难于量化有质的差别的事物(事件)给予量化。如某人面临多种方案的选择工作时，要考虑地点、工作性质、单位福利等。可将要考虑的因素都折合为效用值。得到各方案的综合效用值，然后选择效用值最大的方案，这就是最大效用值决策准则。

在风险情况下，只作一次决策时，再用最大期望值决策准则，就不那么合理了，如表16-14是各方案及按最大收益期望值的计算结果。表16-14的三个方案的EMV都相同，显然这三个方案并不是等价的。另一方面因 EMV* 给出的是平均意义下的最大，当决策后只实现一次时，用 EMV* 决策准则就不恰当了。这时可用最大效用值决策准则来解决这矛盾。

表 16-14

S_i	E_1	E_2	E_3	E_4	EMV
	p_j				
	0.35	0.35	0.15	0.15	
A	418.3	418.3	−60	−60	275
B	650	−100	650	−100	275
C	483	211.3	480	−267	275

16.5.2 效用曲线的确定

确定效用曲线的基本方法有两种，一种是直接提问法；另一种是对比提问法。

1. 直接提问法

直接提问法是向决策者提出一系列问题，要求决策者进行主观衡量并作出回答。例如向某决策者提问："今年你企业获利100万元，你是满意的，那么获利多少，你会加倍满意?"若这决策者回答200万元。这样不断提问与回答，可绘制出这决策者的获利效用曲线。显然这种提问与回答是十分含糊的，很难确切，所以应用较少。

2. 对比提问法

设决策者面临两种可选方案 A_1, A_2。A_1 表示他可无任何风险地得到一笔金额 x_2；A_2 表示他可以概率 p 得到一笔金额 x_1，或以概率$(1-p)$损失金额 x_3；且 $x_1>x_2>x_3$，设 $U(x_1)$表示金额 x_1 的效用值，若在某条件下，这决策者认为 A_1、A_2 两方案等价时，可表示为

$$pU(x_1)+(1-p)U(x_3)=U(x_2) \tag{16-1}$$

确切地讲，这决策者认为 x_2 的效用值等价于 x_1, x_3 的效用期望值。于是可用对比提问法来测定决策者的风险效用曲线。从式(16-1)可见，其中有 x_1、x_2、x_3、p 四个变量，若其中任意三个为已知时，向决策者提问第四个变量应取何值? 并请决策者作出主观判断第四个变量应取的值是多少。提问的方式大致有三种：

(1) 每次固定 x_1、x_2、x_3 的值，改变 p，问决策者："p 取何值时，认为 A_1 与 A_2 等价。"

(2) 每次固定 p、x_1、x_3 的值，改变 x_2，问决策者："x_2 取何值时，认为 A_1 与 A_2 等价。"

(3) 每次固定 p、x_2、x_3(或 x_1)的值，改变 x_3(或 x_1)，问决策者："x_3(或 x_1)取何值时，认为 A_1 与 A_2 等价。"一般采用改进的 V－M(Von Neumann-Morgenstern)法。即每次取 $p=0.5$，固定 x_1、x_3 利用

$$0.5U(x_1)+0.5U(x_3)=U(x_2)$$

改变 x_2 三次，提三问，确定三点，即可绘出这决策者的效用曲线，下面用数字说明。

设 $x_1=1\,000\,000$，$x_3=-500\,000$

取 $U(1\,000\,000)=1$，$U(-500\,000)=0$

$$0.5U(x_1)+0.5U(x_3)=U(x_2) \qquad (16\text{-}2)$$

第一问："你认为 x_2 取何值时，式(16-2)成立?"若回答为"在 $x_2=-250\,000$ 时"，那么 $U(-250\,000)=0.5$，那 x_2 的效用值为 0.5。在坐标系中给出第一个点，见图 16-4。利用

$$0.5U(x_1)+0.5U(x_2)=U(x_2') \qquad (16\text{-}3)$$

提第二问，"你认为 x'_2 取何值时，式(16-3)成立?"若回答为"在 $x'_2=75\,000$ 时"，那么

$$U(75\,000)=0.5\times1+0.5\times0.5=0.75$$

即 x_2'的效用值为 0.75，在坐标系中给出第二个点。利用

$$0.5U(x_2)+0.5U(x_3)=U(x_2'') \qquad (16\text{-}4)$$

提第三问，"你认为 x''_2 取何值时，式(16-4)成立?"若回答为"在 $x''_2=-420\,000$ 时"，那么

$$U(-420\,000)=0.5\times0.5+0.5\times0=0.25$$

即 x''_2 的效用值为 0.25，在坐标系中给出第三点。这就可以绘制出这决策者对风险的效用曲线，见图 16-4。

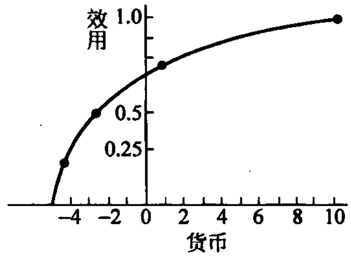


图 16-4

从以上向决策者提问及回答的情况来看，不同的决策者会选择不同的 x_2、x'_2、x''_2 的值，使式(16-2)，式(16-3)，式(16-4)成立。这就能得到不同形状的效用曲线，并表示了不同决策者对待风险的不同态度。一般可分为：保守型、中间型、冒险型三种。其对应的曲线见图 16-5。具有中间型效用曲线的决策者，他认为他的收入金额的增长与效用值的增长成等比关系；具有保守型效用曲线的决策者，他认为他对损失金额愈多愈敏感，相反地对收入的增加比较迟钝，即他不愿承受损失的风险；具有冒险型效用曲线的决策者，他认为他对损失金额比较迟钝，相反地对收入的增加比较敏感，即他可以承受损失的风险。这是三种典型，某一决策者可能兼有三种类型，如图 16-6 所示。

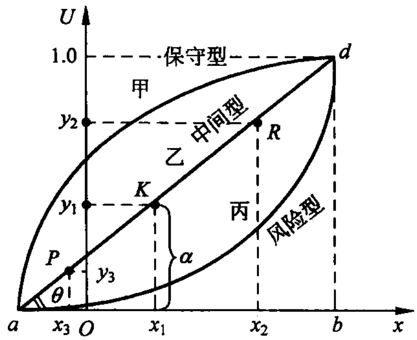


图 16-5

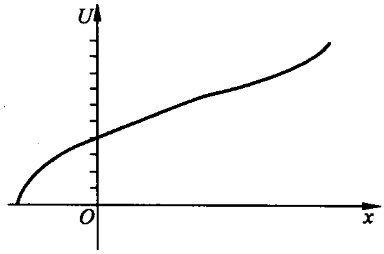


图 16-6

16.5.3 效用曲线的拟合

当用计算机时需用解析式来表示效用曲线,并对决策者测得的数据进行拟合,常用的关系式有以下六种。

(1) 线性函数 $U(x)=c_1+a_1(x-c_2)$

(2) 指数函数 $U(x)=c_1+a_1(1-e^{a_2(x-c_2)})$

(3) 双指数函数 $U(x)=c_1+a_1(2-e^{a_2(x-c_2)}-e^{a_3(x-c_2)})$

(4) 指数加线性函数 $U(x)=c_1+a_1(1-e^{a_2(x-c_2)})+a_3(x-c_2)$

(5) 幂函数 $U(x)=a_1+a_2[c_1(x-a_3)]a_4$

(6) 对数函数 $U(x)=c_1+a_1\log(c_3x-c_2)$

16.6 决策树

有些决策问题,当进行决策后又产生一些新情况,并需要进行新的决策,接着又有一些新情况,又需要进行新的决策。这样决策、情况、决策……构成一个序列,这就是序列决策。描述序列决策的有力工具之一是决策树,决策树是由决策点,事件点及结果构成的树形图。一般选用最大收益期望值和最大效用期望值或最大效用值为决策准则,下面用例子说明。

例 4 设有某石油钻探队,在一片估计能出油的荒田钻探。可以先做地震试验,然后决定钻井与否。或不做地震试验,只凭经验决定钻井与否。做地震试验的费用每次 3 000 元,钻井费用为 10 000 元。若钻井后出油,这井队可收入 40 000 元,若不出油就没有任何收入。各种情况下估计出油的概率已估计出,并标在图 16-7 上。问钻井队的决策者如何作出决策使收入的期望值为最大。

解 上述决策问题用决策树来求解,并将有关数据标在图上,见图 16-7。

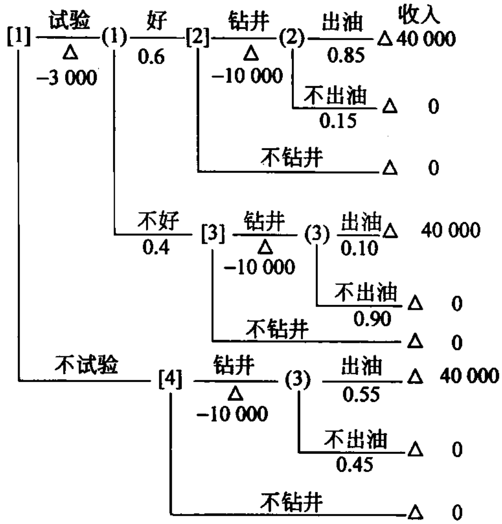


图 16-7

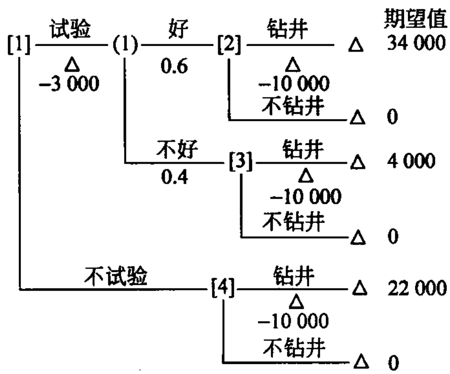
[ ]表示决策点；( )表示事件点；△表示收益点；负值表示支付。

图 16-7 表明这是两级随机决策问题，采用逆决策顺序方法求解，计算步骤是：

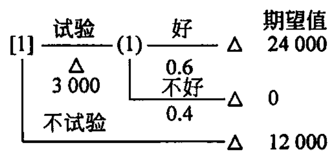
(1) 计算各事件点的收入期望值

事件点	收入期望值
(2)	40 000×0.85+0×0.15=34 000
(3)	40 000×0.10+0×0.90=4 000
(4)	40 000×0.55+0×0.45=22 000

将收入期望值标在相应的各点处，这时可将原决策树(图 16-7)简化为图 16-8(a)。



(a)



(b)

图 16-8

(2) 按最大收入期望值决策准则在图 16-8(a)上给出各决策点的抉择。在决策点[2],按

$$\max[(34\,000-10\,000),0]=24\,000$$

所对应的策略为应选策略,即钻井,在决策点[3],按

$$\max[(4\,000-10\,000),0]=0$$

所对应的策略为应选策,即不钻井,在决策点[4],按

$$\max[(22\,000-10\,000),0]=12\,000$$

所对应的策略为应选策略,即钻井。

(3) 在决策树上保留各决策点的应选方案,把淘汰策略去掉,得到图 16-8(b),这时再计算事件点(1)的收入期望值。

$$24\,000\times0.60+0\times0.40=14\,400$$

将它标在(1)旁。

(4) 决策点[1]有两个方案:做地震试验和不做地震试验,各自的收入期望值为(14 400−3 000)和 12 000。按

$$\max[(14\,400-3\,000),12\,000]=12\,000$$

所对应的策略为应选策略,即不做地震试验。

这个决策问题的决策序列为:选择不做地震试验,直接判断钻井,收入期望值为 12 000 元。

例 5 设决策者的效用曲线如图 16-9 所示。试以最大效用期望值为决策准则,对例 4 进行决策。

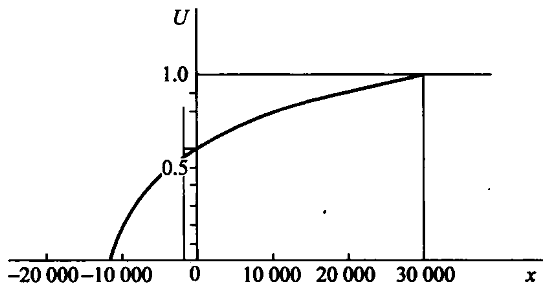


图 16-9

解 如同例 4 一样采用决策树为工具,在决策树的右端标上纯收入

$$纯收入=收入-支出$$

然后由决策者的效用曲线查得各纯收入相应的效用值,并将此值记在相应的纯收入旁,见图 16-10。

决策分析可在图上进行。见图 16-11。以下按逆序先计算事件点(2)、(3)、(4)的效用期望值分别为 0.833、0.098、0.672,并标在相应各点旁,然后在各决策点[2]、[3]、[4]进行选择,其计算为

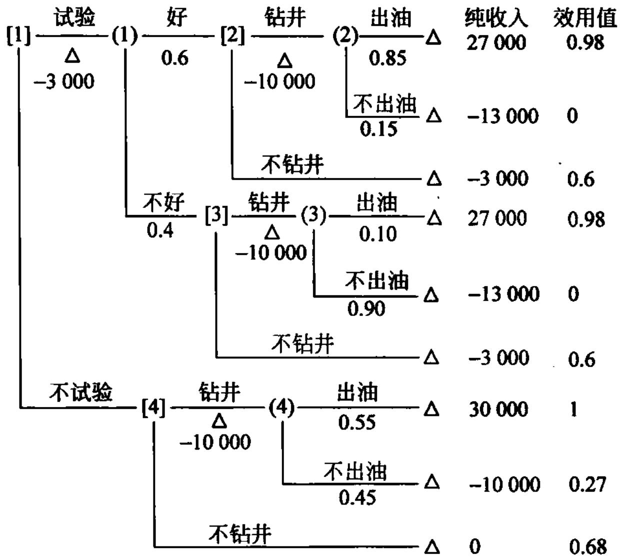


图　16-10

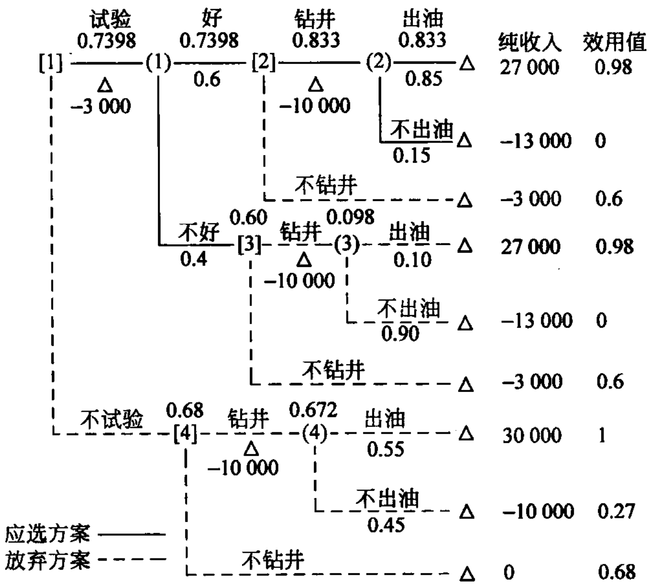


图　16-11

$$\max_2(0.833, 0.60) = 0.833$$

$$\max_3(0.098, 0.60) = 0.60$$

$$\max_4(0.672, 0.68) = 0.68$$

接着计算事件点(1)的效用期望值为 0.739 8，记在(1)旁，决策点[1]的选择为

$$\max(0.7398, 0.68)=0.7398$$

根据以上计算在决策树上可见决策序列为：先做地震试验，若结果好，则钻井；若结果不好，则不钻井。显然这决策是保守型的。因决策者的效用曲线是保守型的。

16.7 灵敏度分析

16.7.1 灵敏度分析的意义

通常在决策模型中自然状态的概率和损益值往往由估计或预测得到，不可能十分正确，此外实际情况也在不断地变化，现需分析为决策所用的数据可在多大范围内变动，原最优决策方案继续有效，进行这种分析称为灵敏度分析，下面用例来说明。

例 6 假设有外表完全相同的木盒 100 只，将其分为两组，一组内装白球，有 70 盒，另一组内装黑球，有 30 盒。现从这 100 盒中任取一盒，请你猜，如这盒内装的是白球，猜对了得 500 分，猜错了罚 150 分；如这盒内装的是黑球，猜对了得 1 000 分，猜错了罚 200 分。为使期望得分最多，应选那一方案，有关数据列于表 16-15。

表 16-15

方案	自然状态(概率)	
	白 0.7	黑 0.3
猜白	500	−200
猜黑	−150	1 000

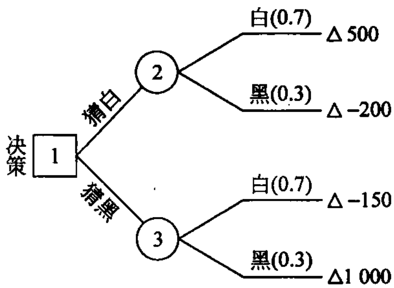


图 16-12

解 先画出决策树，见图 16-12，计算各方案的期望值。

"猜白"的期望值为

$$0.7\times500+0.3\times(-200)=290$$

"猜黑"的期望值为

$$0.7\times(-150)+0.3\times1000=195$$

经比较可知"猜白"方案是最优，现假定出现白球的概率从 0.7 变为 0.8，这时各方案的期望值为

"猜白"的期望值为　$0.8\times500+0.2\times(-200)=360$

"猜黑"的期望值为　$0.8\times(-150)+0.2\times1000=80$

可见猜白方案仍是最优。再假定出现白球的概率从 0.7 变为 0.6，这时各方案的期望值为

"猜白"的期望值为　$0.6\times500+0.4\times(-200)=220$

“猜黑”的期望值为 $0.6\times(-150)+0.4\times1\,000=310$

现在的最优方案不是猜白，而是猜黑了。可见由于各自然状态发生的概率的变化，可引起最优方案的改变。那么转折点如何确定？见16.7.2小节的分析。

16.7.2 转折概率

设 p 为出现白球的概率，$(1-p)$ 为出现黑球的概率。当这两个方案的期望值相等时，即

$$p\times500+(1-p)\times(-200)=p\times(-150)+(1-p)\times1\,000$$

求得 $p=0.648\,6$，称它为转折概率。即当 $p>0.648\,6$，猜白是最优方案；当 $p<0.648\,6$，猜黑是最优方案。p 可推导，表示为

$$p=\frac{a_{12}-a_{22}}{a_{12}-a_{22}+a_{21}-a_{11}}$$

若这些数据在某允许范围内变动，而最优方案保持不变，这方案就是比较稳定的。反之，这些数据在某允许范围内稍加变动，则最优方案就有变化，这方案就是不稳定的。由此可以得出那些非常敏感的变量，那些不太敏感的变量，以及最优方案不变条件下，这些变量允许变化的范围。

习题

16.1 某厂考虑生产甲、乙两种产品，根据过去市场需求统计见表16-16。

表 16-16

方案	自然状态	
	旺　季	淡　季
	概率 $a_1=0.7$	概率 $a_2=0.3$
甲　种	4	3
乙　种	7	2

用最大可能性法进行决策。

16.2 对第16.1题用期望值法进行决策并进行灵敏度分析，求出转折概率。

16.3 某公司为了扩大市场，要举行一个展销会，会址打算选择在甲、乙、丙三地。获利情况除了与会址有关系外，还与天气有关，天气可区分为晴、普通、多雨三种（分别以 N_1、N_2、N_3 表示）。通过天气预报，估计三种天气情况可能发生的概率为0.25，0.50，0.25。其收益情况见表16-17，用期望值准则进行决策。

表 16-17 万元

选址方案	自然状态		
	晴(N_1)	普通(N_2)	多雨(N_3)
	概率 0.25	概率 0.50	概率 0.25
甲 地	4	6	1
乙 地	5	4	1.5
丙 地	6	2	1.2

16.4 将第 16.3 题用矩阵法进行决策。

16.5 今要建立一个企业，有 4 个投资方案，三种自然状态，投资数量见表 16-18。

表 16-18 百万元

投资方案	自然状态		
	Q_1	Q_2	Q_3
	概率 1/2	概率 1/3	概率 1/6
A_1	4	7	4
A_2	5	2	3
A_3	8	6	10
A_4	3	1	9

用矩阵法进行决策。

16.6 某公司需要决定建大厂还是建小厂来生产一种新产品，该产品的市场寿命为 10 年，建大工厂的投资费用为 280 万元，建小厂的投资为 140 万元。10 年内销售状况的离散分布的状态如下：

高需求量的可能性为 0.5；

中等需求量的可能性为 0.3；

低需求量的可能性为 0.2。

公司进行了成本—产量—利润分析，在工厂规模和市场容量的组合下，它们的条件收益如下：

① 大工厂，高需求，每年获利 100 万元。

② 大工厂，中等需求，每年获利 60 万元。

③ 大工厂，低需求，由于开工不足，引起亏损 20 万元。

④ 小工厂，高需求，每年获利 25 万元(供不应求引起销售损失较大)。

⑤ 小工厂，中需求，每年获利 45 万元(销售损失引起的费用较低)。

⑥ 小工厂，低需求，每年获利 55 万元(因工厂规模与市场容量配合得好)。

用决策树方法进行决策。

16.7 将第 16.5 题用决策树法进行决策。

16.8 将第 16.6 题用矩阵法决策。

16.9 将第 16.3 题用决策树方法进行决策。

16.10 某厂有一种新产品，其推销策略有 S_1，S_2，S_3 三种可供选择，但各方案所需的资金、时间都不同，加上市场情况的差别，因而获利和亏损情况不同。而市场情况也有三种：Q_1（需要量大），Q_2（需要量一般），Q_3（需要量低）。市场情况的概率并不知道，其益损矩阵见表 16-19，用乐观法进行决策。

表 16-19 新产品推销益损矩阵表 万元

S_i	市场情况		
	Q_1	Q_2	Q_3
S_1	50	10	−5
S_2	30	25	0
S_3	10	10	10

16.11 将 16.10 题用悲观准则进行决策。

16.12 某企业面临三种方案可以选择，五年内的益损表见表 16-20，用乐观系数法（$a_1=0.3$，$a_2=0.7$）决策，然后加以比较。

表 16-20 企业改建益损表

决策方案	自然状态 需求量			
	高	中	低	失败
扩建	50	25	−25	−45
新建	70	30	−40	−80
转包	30	15	−1	−10

16.13 将 16.12 题用等可能准则（Laplace）进行决策，并与上题比较结果。

16.14 在开采油井时，出现不定情况，用后悔值准则决定是否开采。益损矩阵见表 16-21。

表 16-21 油井开采益损矩阵表

方案	自然状态	
	有 油	无 油
	Q	$\bar{Q}$
开 采	5	1
不开采	0	0

16.15 某企业要投产一种新产品，投资方案有三个：S_1、S_2、S_3，不同经济形势下的利润如表16-22所示。用乐观系数准则($a_1=0.6, a_2=0.4$)进行决策。

表 16-22 新产品投资利润表 万元

投资方案	不同经济形势		
	好	平	差
S_1	10	0	−1
S_2	25	10	5
S_3	50	0	−40

16.16 将16.15题用等可能准则进行决策。

16.17 建厂投资有三个行动方案可以选择，并有三种自然状态，其损失表见表16-23，用乐观准则进行决策。

表 16-23 建厂投资损失表

方 案	自然状态		
	Q_1	Q_2	Q_3
A_1	3	7	3
A_2	6	5	4
A_3	5	6	10

16.18 将16.17题用悲观准则进行决策。

参考资料

[1] 姜青舫编著. 实用决策分析. 贵阳：贵州人民出版社，1985
[2] 黄孟藩. 管理决策概论. 北京：中国人民大学出版社，1982
[3] Milan Zeleny. Multiple Criteria Decision Making. McGraw Hill Book Company, 1982

CHAPTER 17
第17章

多目标决策

17.1 引言

在生产、经济、科学和工程活动中经常需要对多个目标(指标)的方案、计划、设计进行好坏的判断,例如设计一个导弹,既要其射程远,又要耗燃料少,还要命中率高等;又如选择新厂的厂址,除了要考虑运费、造价、燃料供应费等经济指标外,还要考虑对环境的污染等社会因素,只有对各种因素的指标进行综合衡量后,才能作出合理的决策。

例1 由 n 种成分 $x_1,x_2,\cdots,x_n$ 组成一个橡胶配方,可用 $x=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$ 表示。对于每一个配方要同时考查几个指标,如强度 f_1,硬度 f_2,伸长率 f_3,变形度 f_4 等。假定有 m 个指标,它们都与配方方案 x 有关,它们与 x 的关系为 $f_1(x),f_2(x),\cdots,f_m(x)$。当 m 很多时,要比较两方案的优劣时,就往往很难下决断了。于是有人把这问题用数学规划来处理。先以某指标作为主要指标,如以强度 f_1 为主要指标,并且越大越好。而其他指标只要落在一定规格范围内就可以。这就把这问题化为求

$$\max_{x\in R} f_1(x)$$

$$R=\{x \mid f_i'\leqslant f_i(x)\leqslant f_i'',i=2,\cdots,m,x\in A\}$$

这里 A 表示对 x 本身的一个限制,f_i'、f_i'' 分别表示第 i 个指标的上、下限。

例2 门坐式起重机四连杆变幅机构的最优设计。

衡量变幅机构设计的好坏一般要求:

(1) 货物尽量走水平,即高度变化小;

(2) 货物在移动时,水平速度的变化量尽量小;

(3) 由吊装的货物所引起的倾覆力矩尽量小。

为了达到这三项要求,需用五个指标 $f_1,\cdots,f_5$ 来描述。关于要求(1),用落差 $f_1=\Delta y$ 表示;要求(2),用速比 $f_2=i_v$ 表示;要求(3),用三个指标:大幅度时力矩 $f_3=M(a_0)$;小幅度时力矩 $f_4=M(a_1)$,力矩变化量 $f_5=\Delta M$。要求 f_1、f_2、f_5 越小越好;f_3、f_4 适中为好。

对这样一个多目标问题可采用功效系数法来解决。可以列举出很多种多目标问题,针对不同的问题可采用不同的方法去解决。

17.2 基本概念

在考虑单目标最优化问题时，只要比较任意两个解对应的目标函数值后，就能确定谁优谁劣（目标值相等时除外）。在多目标情况下，就不能作这样简单的比较来确定谁优谁劣了。例如有两个目标都要求实现最大化，这样的决策问题，若能列出10个方案，各方案能实现的不同的目标值如图17-1所示。从图中可见，对于第一个目标来讲方案①优于②；而对于第二个目标则方案②优于①。因此无法确定谁优谁劣；但是它们都比方案③、⑤劣。方案③、⑤之间又无法相比。在图17-1中10个方案，除方案③、④、⑤以外，其他方案都比它们中的某一个劣。因而称方案①、②、⑥、⑦、⑧、⑨、⑩为劣解，而方案③、④、⑤之间又无法比较谁优谁劣，但又不存在一个比它们中任一个还好的方案，故称此三个方案为非劣解（或称为有效解）。以后再给出确切的定义。

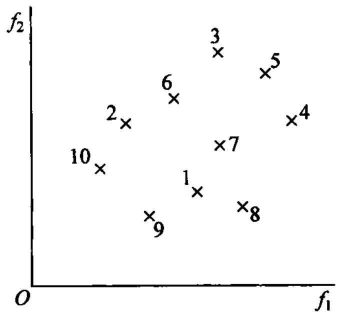


图 17-1

假定有 m 个目标 $f_1(x),\cdots,f_m(x)$ 同时要考查，并要求都越大越好。在不考虑其他目标时，记第 i 个目标的最优值为

$$f_i^{(o)}=\max_{x\in R} f_i(x)$$

相应的最优解记为 $x^{(i)}, i=1,2,\cdots,m$；其中 R 是解的约束集合。

$$R=\{x\,|\,g(x)\geqslant 0\}\qquad g(x)=\{g_1(x),\cdots,g_l(x)\}^{\mathrm{T}}$$

当这些 $x^{(i)}$ 都相同时，就以这共同解作为多目标的共同最优解。一般不会全相同，例如 $x^{(1)}\neq x^{(2)}$ 时，这两个解就难比优劣，但是它们一定都是非劣解。为了与单目标最优化的记号有所区别，今后用

$$V\text{——}\max_{x\in R} F(x)\quad 或\quad V\text{——}\max_{g(x)\geqslant 0} F(x)$$

表示在约束集合 R 内求多目标问题的最优（亦称求向量最优）；其中

$$F(x)=\{f_1(x),\cdots,f_m(x)\}^{\mathrm{T}}$$

若各目标值都要求越小越好，就用下式表示。

$$V\text{——}\min_{x\in R} F(x)$$

下面考查使目标值越大越好，为了简易起见，本节一般只考虑 n 维欧氏空间 E^n，即

$$x=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}\in E^n,\quad R\subset E^n,\quad F(x)\in E^m$$

实际上当 x_0 是最优解时，即表示 $\forall x\in R$，有

$$F(x)\leqslant F(x_0)$$

当 x_0 是非劣解时，即不存在 $x\in R$，有

$$F(x) \geqslant F(x_0)$$

以后用"≥"表示 $F(x) \geqq F(x_0)$，但 $F(x) \neq F(x_0)$，即至少有一个分量，">"才成立，即一定大于。相应的 $F(x_0)$ 在目标函数空间中称为非劣点或有效点。有的还进一步引入弱非劣解，即当 x_0 是弱非劣解，若不存在 $x \in R$，有

$$F(x) > F(x_0)$$

为直观起见，举几个数值例子。

例 3 设 $f_1(x)=2x-x^2$，$f_2(x)=\begin{cases} x & 0 \leqslant x \leqslant 1 \\ -2x+3, & 1 \leqslant x \leqslant 2 \end{cases}$

$R=[0,2]$，求 $V\text{——}\max\limits_{x \in R} F(x)$

解 先对单个目标分别求出其最优解。显然第一个目标的最优解 $x^{(1)}=1$，这时

$$f_1^{(0)}=f_1(x=1)=\max_{x \in R} f_1(x)$$

第二个目标的最优解是 $x^{(2)}=1$，这时

$$f_2^{(0)}=f_2(x=1)=\max_{x \in R} f_2(x)$$

因为 $x^{(1)}=x^{(2)}=1$，故取 $x^*=1$ 作为这多目标问题的最优解。下面用变量空间和目标函数空间分别来描述各种解的情况，见图 17-2。图中 α 和 β 两个解彼此无法比较，但都劣于 $x^*=1$。

例 4 设 $f_1(x)=2x-x^2$，$f_2(x)=x$，$R=[0,2]$，求 $V\text{——}\max\limits_{x \in R} F(x)$

解 容易求得 $x^{(1)}=1$，$x^{(2)}=2$，这时多目标问题没有共同最优解，从图 17-3 可见，α 和 β 两个解彼此无法比较，但是容易找到 α' 比 α 优，α' 比 β 仍无法比较优劣，但还可找到 α'' 比 α' 优。解 β 却不存在 β' 可以比它优，这时 β 为非劣解。本例中 $x \in [1,2]$ 时都是非劣解。

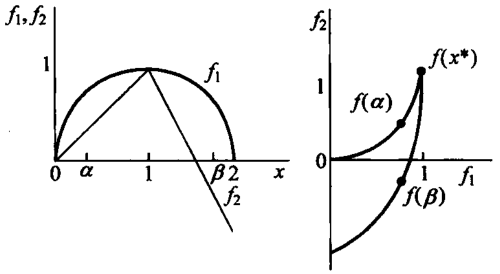


图 17-2

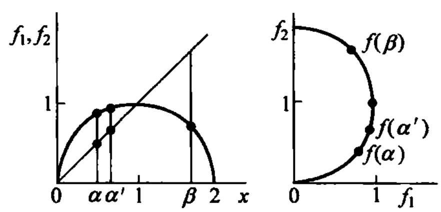


图 17-3

例 5 设 $f_1(x)=2x-x^2$，$f_2(x)=\dfrac{1}{8}(-12x^2+36x-15)$，$R=[0,\ 2]$，求

$$V\text{——}\max_{x \in R} F(x)$$

解 易求得 $x^{(1)}=1$，$x^{(2)}=1.5$，这时多目标问题无最优解，而 $x \in [1,1.5]$ 都是非劣解，见图 17-4。

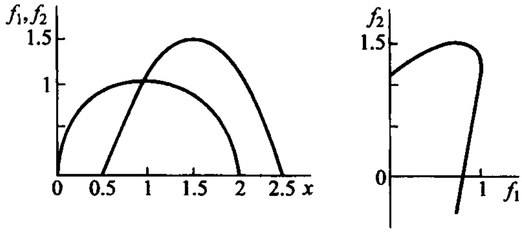


图 17-4

例 6 设 $f_1(x)=-3x_1+2x_2$，$f_2(x)=x_1+2x_2$

$$R:\begin{cases}-2x_1-3x_2+18\geqslant 0\\ -2x_1-x_2+10\geqslant 0\\ x_1\geqslant 0,\ x_2\geqslant 0\end{cases}$$

求 $V\text{——}\max\limits_{x\in R}F(x)$

解 易求得 $x^{(1)}=(0,6)$，$x^{(2)}=(0,6)$，因而多目标问题最优解即 $x^*=(0,6)$。图 17-5所示的 α、β 之间无法比较，但都劣于 A。

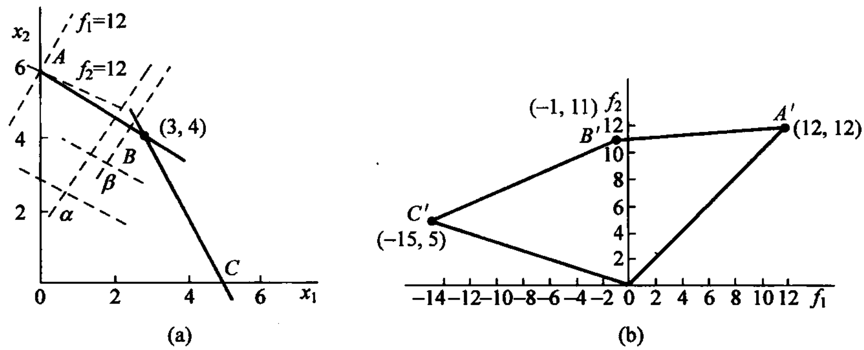


图 17-5

例 7 设 $f_1(x)=-3x_1+2x_2$，$f_2(x)=4x_1+3x_2$，R 同例 6，求 $V\text{——}\max\limits_{x\in R}F(x)$

解 易求得 $x^{(1)}=(0,6)$，$x^{(2)}=(3,4)$，这时多目标问题无最优解，但有非劣解，在联结点 $x^{(1)}=(0,6)$ 到 $x^{(2)}=(3,4)$ 之间的线段上的点(见图 17-6(a))都是非劣解，而连接点 A'、B'(见图 17-6(b))的直线上的点都是非劣点。

在单目标时任何两个解都可以比较其优劣，因此是完全有序的。可是在多目标时，任何两个解不一定都是可比出其优劣的，因此只能是半有序的。假定所有解 x 是属于全空间 $\sum$ 中某一个约束集合 R，即 $x\in R\subset\sum$，在 $\sum$ 上对任一个解 x 可以定义一个半序：$\precsim$，($a\succ b$ 表示 a 优于 b)，且可把 $\sum$ 分成三个子集：

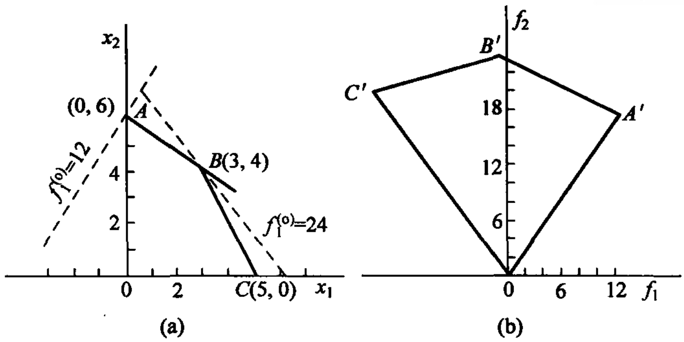


图 17-6

(1) $\sum_{>}(x)$ 所有比 x 优的解集合;

(2) $\sum_{\preccurlyeq}(x)$ 所有比 x 劣或相等的解集合;

(3) $\sum_{\sim}(x)$ 所有与 x 无法比较的解集合。

显然
$$\sum = \sum_{>}(x) \cup \sum_{\preccurlyeq}(x) \cup \sum_{\sim}(x)$$

按照这些子集的划分,Zadeh 给出“非劣”和“最优”的定义①

定义 1 解 $x_0 \in R \subset \sum$ 叫做在 R 内“非劣”,如果 $R \cap \sum_{>}(x_0) = \varnothing$。

定义 2 解 $x_0 \in R \subset \sum$ 叫做在 R 内最优,如果 $R \subset \sum_{\preccurlyeq}(x_0)$。

推论 若 x_0 是最优解,则必为非劣解。反之不然。

证 因为由定义 2,x_0 是最优,$R \subset \sum_{\preccurlyeq}(x_0)$,而 $\sum_{\preccurlyeq}(x_0) \cap \sum_{>}(x_0) = \varnothing$,故 $R \cap \sum_{>}(x_0) = \varnothing$,所以 x_0 也为非劣解。

这推论反之不一定对,即 x_0 是非劣解,有 $R \cap \sum_{>}(x_0) = \varnothing$ 成立。但是由于 R 不一定全在 $\sum_{\preccurlyeq}(x_0)$ 内,而只能保证 $R \subset \sum_{\preccurlyeq}(x_0) \cup \sum_{\sim}(x_0)$,所以不一定是最优解。但在单目标时,由于是全有序,因此 $\sum_{\sim}(x_0) = \varnothing$,所以 x_0 是非劣时,必有 $R \subset \sum_{\preccurlyeq}(x_0)$,即 x_0 自然就是最优解。由此可见在单目标最优化问题时,对最优和非劣可以不区分;但在多目标最优化问题时,这两个概念必须加以区别。

17.3 化多为少的方法

要求若干目标同时都实现最优往往是很难的。经常是有所失才能有所得,那么问题的失得在何时最好。各种不同的思路可引出各种合理处理得失的方法。以下介绍化多为

① 这里用“非劣解”的名词,主要来源于 Zadeh。有不少文章使用“有效解”,还有人用“Pareto 解”等各种名词,其含义基本一致。

少的方法。

由于直接解决多目标问题的最优化较困难，于是很多人想办法将它们化为较容易求解的单目标或双目标问题。由于化法不一，就形成多种方法。

17.3.1 主要目标法

解决主要问题，并适当兼顾其他要求。

1. 优选法

在实际问题中通过分析讨论，抓住其中一两个主要目标，让它们尽可能地好，而其他指标只要满足一定要求即可，通过若干次试验以达到最佳。

2. 数学规划法

设有 m 个目标 $f_1(x), f_2(x), \cdots, f_m(x)$ 要考查，其中方案变量 $x \in R$（约束集合），若以某目标为主要目标，如 $f_1(x)$ 要求实现最优（最大），而对其他目标只满足一定规格要求即可，如

$$f_i' \leqslant f_i(x) \leqslant f_i'' \quad (i=2,\cdots,m)$$

其中当

$f_i' = -\infty$ 或 $f_i'' = \infty$ 就变成单边限制，这样问题便可化成求下述非线性规划问题：

$$\max_{x \in R'} f_1(x)$$

$$R' = \{x \mid f_i' \leqslant f_i(x) \leqslant f_i'', i=2,\cdots,m, x \in R\}$$

17.3.2 线性加权和法

若有 m 个目标 $f_i(x)$，分别给以权系数 $\lambda_i (i=1,2,\cdots,m)$，然后作新的目标函数（也称效用函数）。

$$U(x) = \sum_{i=1}^{m} \lambda_i f_i(x) \text{ ①}$$

该方法的难点是如何找到合理的权系数，使多个目标用同一尺度统一起来。同时所找到的最优解又是向量极值的好的非劣解。在多目标最优化问题中不论用何方法，至少应找到一个非劣解（或近似非劣解）。其次，因非劣解可能有很多，如何从中挑出较好的解，这个解有时就要用到另一个目标。下面介绍几种选择特定权系数的方法。

1. α—法

先以两个目标为例，假设一个目标是要求劳动量消耗 $f_1(x)$ 为最小，另一个目标是收益 $f_2(x)$ 为最高。它们都是线性函数，都以元为单位。R 也为线性约束，即

---

① 一般用和形式的新目标函数，其中所有目标必须是具有相同量纲，若量纲不同，必须进行统一量纲或无量纲化处理，以后无特别说明，一般都指经处理过的目标。

$$R=\{x \mid Ax \leqslant b\}$$

A 为矩阵，b 为列向量。

作为新目标函数

$$U(x)=\alpha_2 f_2(x)-\alpha_1 f_1(x)$$

其中 α_1, α_2 由下述方程组来确定

$$-\alpha_1 f_1^0+\alpha_2 f_2^*=c_1$$

$$-\alpha_1 f_1^*+\alpha_2 f_2^0=c_1$$

其中

$$f_1^0=\min_{x\in R} f_1(x)=f_1(x^{(1)}), f_2^*=f_2(x^{(1)})①$$

$$f_2^0=\max_{x\in R} f_2(x)=f_2(x^{(2)}), f_1^*=f_1(x^{(2)})$$

c_1 可为任意的常数($c_1\neq 0$)，这方程组可解得

$$\alpha_1=\frac{c_1(f_2^0-f_2^*)}{f_1^* f_2^*-f_1^0 f_2^0}$$

$$\alpha_2=\frac{c_1(f_1^*-f_1^0)}{f_1^* f_2^*-f_1^0 f_2^0}$$

若规定 $\alpha_1+\alpha_2=1$，即可得到

$$c_1=\frac{f_1^* f_2^*-f_1^0 f_2^0}{f_2^0-f_2^*+f_1^*-f_1^0}$$

从而有

$$\alpha_1=\frac{f_2^0-f_2^*}{f_2^0-f_2^*+f_1^*-f_1^0}$$

$$\alpha_2=\frac{f_1^*-f_1^0}{f_2^0-f_2^*+f_1^*-f_1^0}$$

易见

$$\frac{\alpha_1}{\alpha_2}\doteq\frac{f_2^0-f_2^*}{f_1^*-f_1^0}=k$$

由这样定义的 α_1、α_2 作出的新目标函数为

$$U(x)=\alpha_2 f_2(x)-\alpha_1 f_1(x)$$

$$=\frac{(f_1^*-f_1^0)f_2(x)-(f_2^0-f_2^*)f_1(x)}{f_2^0-f_2^*+f_1^*-f_1^0}$$

当要求实现最大时，可表示为

$$\max_{x\in R} U(x)=\max_{x\in R}(\alpha_2 f_2(x)-\alpha_1 f_1(x))$$

---

① 假定 $x^{(1)}$ 是唯一的。若有多个 $x^{(1)}$ 达到最小，就必须从中进一步找出使 f_2 达到最优的那个解，同样 $x^{(2)}$ 也应如此处理。

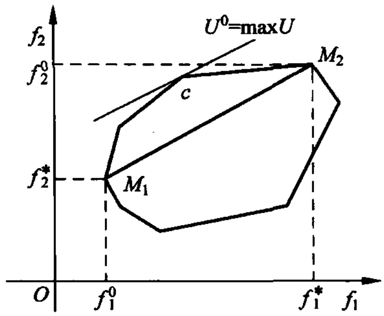


图 17-7

若作目标值空间(f_1,f_2)，则 $U(x)$ 取不同数时，相当一簇平行线，其斜率为 $k=\alpha_1/\alpha_2$。

请注意点 $M_1(f_1^0,f_2^*)$ 与 $M_2(f_1^*,f_2^0)$ 的连线的斜率为 $(f_2^0-f_2^*)/(f_1^*-f_1^0)$，与新目标函数 $U(x)$ 的平行线簇的斜率是一致的，见图 17-7。U 取最大值时，正好是此平行线簇中与 c 点相交。

对于有 m 个目标 $f_1(x),\cdots,f_m(x)$ 的情况，不妨设其中 $f_1(x),\cdots,f_k(x)$ 要求最小化，而 $f_{k+1}(x),\cdots,f_m(x)$ 要求最大化，这时可构成下述新目标函数。

$$\max_{x\in R}U(x)=\max_{x\in R}\left\{-\sum_{j=1}^{k}\alpha_j f_j(x)+\sum_{j=k+1}^{m}\alpha_j f_j(x)\right\}$$

其中$\{\alpha_j\}$满足方程组

$$-\sum_{j=1}^{k}\alpha_j f_{ij}+\sum_{j=k+1}^{m}\alpha_j f_{ij}=c_1,\quad i=1,\cdots,m$$

其中
$$f_{ii}=f_i^0=\min_{x\in R}f_i(x)=f_i(x^{(i)}),\quad i=1,2,\cdots,k$$
$$f_{ii}=f_i^0=\max_{x\in R}f_i(x)=f_i(x^{(i)}),\quad i=k+1,\cdots,m$$
$$f_{ij}=f_j(x^{(i)}),\quad j\neq i,\quad i,j=1,2,\cdots,m$$

例 8 设有 $f_1(x)=4x_1+x_2\to\min$，$f_2(x)=3x_1+2x_2\to\max$

$R=\{x\mid 2x_1+x_2\leqslant 4,x_1+x_2\leqslant 3,x_1,x_2\geqslant 0,x\in E^2\}$，试用 α—法求解。

解 先分别对 $f_1(x),f_2(x)$ 求得其最优解，它们是

$$f_1(x^{(1)})=f_1(0,0)=\min_{x\in R}f_1(x)=f_1^0=0$$
$$f_2(x^{(2)})=f_2(1,2)=\max_{x\in R}f_2(x)=f_2^0=7$$

然后求出
$$f_2^*=f_2(x^{(1)})=0,f_1^*=f_1(x^{(2)})=6$$

由此可得

$$\alpha_1=\frac{f_2^0-f_2^*}{f_2^0-f_2^*+f_1^*-f_1^0}=\frac{7}{7+6}=\frac{7}{13},\alpha_2=\frac{6}{13}$$

$$U(x)=-\alpha_1 f_1(x)+\alpha_2 f_2(x)=-\frac{7}{13}f_1(x)+\frac{6}{13}f_2(x)=\frac{1}{13}(-10x_1+5x_2)$$

易求得
$$\max_{x\in R}U(x)=U(0,3)=15/13$$

2. λ—法

当 m 个目标都要求实现最大时，可用下述加权和效用函数，即

$$U(x)=\sum_{i=1}^{m}\lambda_i f_i(x)$$

其中 λ_i 取
$$\lambda_i=1/f_i^0,\quad f_i^0=\max_{x\in R}f_i(x)$$

17.3.3 平方和加权法

设有 m 个规定值 $f_1^*,\cdots,f_m^*$，要求 m 个函数 $f_1(x),\cdots,f_m(x)$ 分别与规定的值相差尽量小，若对其中不同值的要求相差程度又可不完全一样，即有的要求重一些，有的轻一些。这时可采用下述评价函数：

$$U(x)=\sum_{i=1}^{m}\lambda_i[f_i(x)-f_i^*]^2$$

要求 $\min\limits_{x\in R}U(x)$，其中 λ_i 可按要求相差程度分别给出。

17.3.4 理想点法

有 m 个目标 $f_1(x),\cdots,f_m(x)$，每个目标分别有其最优值

$$f_i^0=\max_{x\in R}f_i(x)=f_i(x^{(i)})\qquad i=1,2,\cdots,m$$

若所有 $x^{(i)}(i=1,2,\cdots,m)$ 都相同，设为 x^0，则令 $x=x^0$ 时，对每个目标都能达到其各自的最优点，可惜一般做不到，因此对向量函数

$$\boldsymbol{F}(x)=(f_1(x),\cdots,f_m(x))^{\mathrm{T}}$$

来说，向量

$$\boldsymbol{F}^0=(f_1^0,\cdots,f_m^0)^{\mathrm{T}}$$

只是一个理想点(即一般达不到它)。

Салуквадзе 提出的理想点法，其中心思想是定义了一定的模，在这个模意义下找一个点尽量接近理想点，即让模

$$\|\boldsymbol{F}(x)-\boldsymbol{F}^0\|\to\min\|\boldsymbol{F}(x)-\boldsymbol{F}^0\|$$

对于不同的模，可以找到不同意义下的最优点，这个模也可看做评价函数，一般定义的 p-模是：

$$\|\boldsymbol{F}(x)-\boldsymbol{F}^0\|=\left[\sum_{i=1}^{m}(f_i^0-f_i(x))^p\right]^{1/p}=L_p(x)$$

其中，p 的取值一般在$[1,\infty)$。当取 $p=2$，这时模即为欧氏空间中向量 $\boldsymbol{F}(x)$ 与向量 $\boldsymbol{F}^0$ 的距离，见图 17-8。要求模最小，也就要找到一解，它对应的目标值与理想点的目标值距离最近，可表示为

$$\min_{x\in R}L_p(x)$$

当 $p=1$ 时，$L_1(x)=\sum_{i=1}^{m}(f_i^0-f_i(x))$

当 $p=\infty$时，$L_\infty(x)=\max\limits_{1\leqslant i\leqslant m}|f_i^0-f_i(x)|$

如 $x_1=(8,6)$，$x_2=(4,3)$两点之间，当 $p=1,2,\cdots,\infty$时其距离的取值见表 17-1。

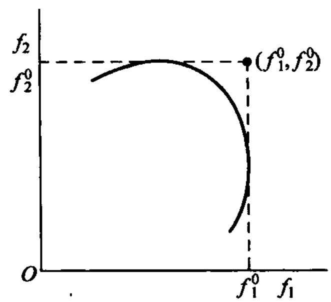


图 17-8

表 17-1

p	$(x_1^1-x_1^2)^p$	$(x_2^1-x_2^2)^p$	$L_p(x)$
1	4	3	7
2	16	9	5
3	64	27	4.498
⋮	⋮	⋮	⋮
∞	4^{∞}	3^{∞}	4

当 $p=2$ 时，其几何意义是两点之间的最短距离为直线；而当 $p>2$ 时，其距离就小于这两点之间的直线距离；并且 p 愈大，距离值就愈趋向于较大的分量(属性、目标)。因此可取不同的 p 值代表人们对较大分量(属性、目标)的偏爱程度，它就不是几何概念了。

上述 17.3.3 小节和 17.3.4 小节的方法也是目标规划法的一类，即事先规定一些指标值，然后另设目标，看其接近这些规定值的程度。新设的目标有时也称超目标，易证明理想点法求出的解一定是非劣解，自然它在目标值空间中就是有效点。

例 9 设 $f_1(x)=-3x_1+2x_2$，$f_2(x)=4x_1+3x_2$ 都要求实现最大。约束集为

$$R=\{x\mid 2x_1+3x_2\leqslant 18,2x_1+x_2\leqslant 10,x_1x_2\geqslant 0,x\in E^2\}$$

试用理想点法求解。

解 先分别对单目标求出最优解 $x^{(1)}=(0,6)$，$x^{(2)}=(3,4)$。对应的目标值为

$$f_1(x^{(1)})=f_1(0,6)=f_1^0=12$$

$$f_2(x^{(2)})=f_2(3,4)=f_2^0=24$$

故理想点为

$$F^0=(f_1^0,f_2^0)=(12,24)$$

取 $p=2$，这时要求

$$\min_{x\in R} L_2(x)=\{[f_1(x)-f_1^0]^2+[f_2(x)-f_2^0]^2\}^{1/2}$$

这时可求得最优解为 $x^*=(0.53,5.65)$，对应的目标值分别为 $f_1^*=9.72$，$f_2^*=19.06$，见图 17-9。

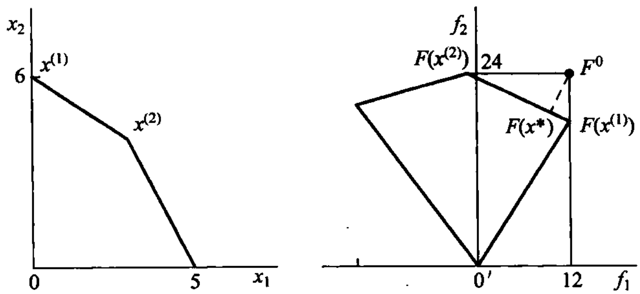


图 17-9

17.3.5 乘除法

当在 m 个目标 $f_1(x),\cdots,f_m(x)$ 中，不妨设其中 k 个 $f_1(x),\cdots,f_k(x)$ 要求实现最小，其余 $f_{k+1}(x),\cdots,f_m(x)$ 要求实现最大，并假定

$$f_{k+1}(x),\cdots,f_m(x)>0$$

这时可采用评价函数

$$U(x)=\frac{f_1(x)f_2(x)\cdots f_k(x)}{f_{k+1}(x)\cdots f_m(x)}\rightarrow\min$$

17.3.6 功效系数法——几何平均法

设 m 个目标 $f_1(x),\cdots,f_m(x)$，其中 k_1 个目标要求实现最大，k_2 个目标要求实现最小，其余的目标是过大不行，过小也不行。对于这些目标 $f_i(x)$ 分别给以一定的功效系数（即评分）d_i，d_i 是在[0,1]之间的某一数。当目标最满意达到时，取 $d_i=1$；当最差时取 $d_i=0$。描述 d_i 与 $f_i(x)$ 的关系，称为功效函数，表示为 $d_i=F_i(f_i)$。对于不同类型目标应选用不同类型的功效函数。

Ⅰ型：当 f_i 越大；d_i 也越大；f_i 越小，d_i 也越小。

Ⅱ型：f_i 越小，d_i 越大；f_i 越大，d_i 越小。

Ⅲ型：当 f_i 取适当值时，d_i 最大；而 f_i 取偏值（即过大或过小）时，d_i 变小。

具体功效函数构造法可以很多，有直线法，折线法，见图 17-10 和图 17-11，指数法见图 17-12。

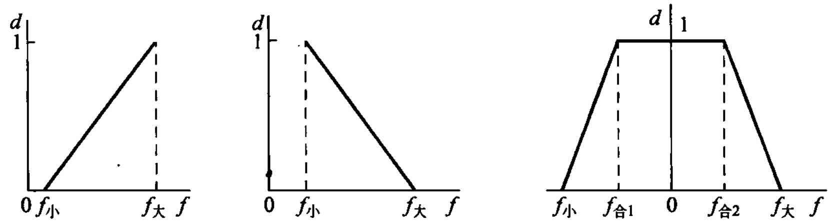


图 17-10

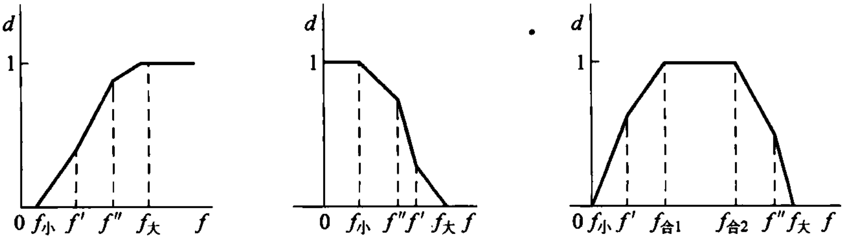


图 17-11

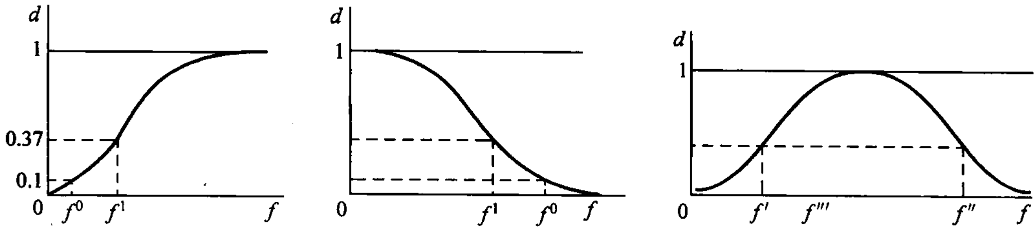


图 17-12

用指数法构造Ⅰ型功效函数(见图 7-12),可设其表达式为

$$d=\mathrm{e}^{-(\mathrm{e}^{-(b_0+b_1 f)})}$$

其中 b_0, b_1 可这样确定:

当 f 达到某一刚合格值 f^1 时,取 $d^1=\mathrm{e}^{-1}\doteq 0.367\,9$

当 f 达到某一不合格值 f^0 时,取 $d^0=\mathrm{e}^{-\mathrm{e}}\doteq 0.065\,98$

将上述要求代入上式即有

$$d^1=\mathrm{e}^{-1}=\mathrm{e}^{-(\mathrm{e}^{-(b_0+b_1 f^1)})}$$

$$d^0=\mathrm{e}^{-\mathrm{e}}=\mathrm{e}^{-(\mathrm{e}^{-(b_0+b_1 f^0)})}$$

由这两式可得

$$b_0+b_1 f^1=0$$
$$b_0+b_1 f^0=-1$$

解之得

$$b_0=\frac{f^1}{f^0-f^1},\quad b_1=\frac{-1}{f^0-f^1}$$

即

$$d=\mathrm{e}^{-\mathrm{e}^{\left(\frac{f-f^1}{f^0-f^1}\right)}}$$

同样对Ⅱ型功效函数,可取为

$$d=1-\mathrm{e}^{-\mathrm{e}^{\left(\frac{f-f'}{f''-f'}\right)}}$$

对于Ⅲ型功效函数,取

$$d=\mathrm{e}^{-|Y|^n}$$

其中

$$Y=\frac{2f-(f'+f'')}{f'-f''}$$

这样,当 $f=f'$ 或 $f=f''$ 时,$Y=\pm 1, d=\mathrm{e}^{-1}$,为刚好可接受的值;当 $f=\dfrac{f'+f''}{2}$ 时,$Y=0$,$d=1$,即 f 达到比较适当的值。为了确定 n,可再取一个 f''',使其与某一个适当的 d 值相

对应，这时可给出

$$n=\frac{\ln(\ln 1/d)}{\ln|Y|}$$

例如，取 $f=f'''=\frac{2f'+f''}{3}$，这时 $Y=1/3$，使其与 $d=e^{-1/2}$ 相对应（见图 17-12），则

$$n=\frac{\ln 1/2}{\ln 1/3}\approx 0.6309$$

即

$$d=e^{-|Y|^{0.6309}}, Y=\frac{2f-(f'+f'')}{f'-f''}$$

有了功效函数后，对每个目标都可对应为相应的功效函数。目标值可转换为功效系数。这样每确定一方案 x 后，就有 m 个目标函数值 $f_1(x),\cdots,f_m(x)$。然后用其对应的功效函数转换为相应的功效系数 $d_1,\cdots,d_m$。并可用它们的几何平均值。

$$D=\sqrt[m]{d_1 d_2 \cdots d_m}$$

为评价函数，显然 D 越大越好，$D=1$ 是最满意的，$D=0$ 是最差的，这样定义的评价函数有一个好处，一个方案中只要有一个目标值太差，如 $d_i=0$，就会使 $D=0$，而不会采用这个方案。

17.4 分层序列法

由于同时处理 m 个目标是比较麻烦，故可采用分层法。分层法的思想是把目标按其重要性给出一个序列，分为最重要目标、次要目标等。设给出的重要性序列为

$$f_1(x), f_2(x), \cdots, f_m(x)$$

下面介绍逐个地求最优化的序列最优化。

首先对第一个目标求最优，并找出所有最优解的集合记为 R_0。然后在 R_0 内求第二个目标的最优解，记这时的最优解集合为 R_1，如此等等一直到求出第 m 个目标的最优解 x^0，其模型如下：

$$f_1(x^0)=\max_{x\in R_0\subset R} f_1(x)$$
$$f_2(x^0)=\max_{x\in R_1\subset R_0} f_2(x)$$
$$\vdots$$
$$f_m(x^0)=\max_{x\in R_{m-1}\subset R_{m-2}} f_m(x)$$

该方法有解的前提是 $R_0,R_1,\cdots,R_{m-1}$ 非空，同时 $R_0,R_1,\cdots,R_{m-2}$ 都不能只有一个元素，否则就很难进行下去。

当 R 是紧致集，函数 $f_1(x)\cdots f_m(x)$ 都是上半连续，则按下式定义的集求解。

$$R_{k-1}^{*}=\{x\mid f_k(x)=\sup_{u\in R_{k-2}^{*}} f_k(u);\ x\in R_{k-2}^{*}\}$$

$k=1,2,\cdots,m$，其中 $R_{k-1}^{*}=R$ 都非空，特别 R_{m-1}^{*} 是非空。故有最优解，而且是共同的最优解。

17.5 直解求非劣解

上述种种方法的基本点是将多目标最优化问题转换为一个或一系列单目标最优化问题。把对后者求得的解作为多目标问题的解，这种解往往是非劣解。对经转换后的问题所求出的最优解往往只是原问题的一个(或部分)非劣解，至于其他非劣解的情况却不得而知。于是出现第三类直接求所有非劣解的方法，当这些非劣解都找到后，就可供决策者作最后的选择，选出的好解就称为选好解。显然决策者这时的选好，必须取决于他心中的另一个目标。这可能是定性的或无法奉告的。运筹学工作者主要是根据已知的目标，尽可能地列出非劣解，以供决策者选择，非劣解求法很多，这里仅介绍线性加权和改变权系数的方法。

在第 3 节中已提到了线性加权和的方法，但那里是按一定想法确定权系数，然后组成线性加权和的函数，并从中求出最优解。可以证明当对目标函数作一定假设，例如目标函数都是严格凹函数，则用线性加权和法求得的最优解是多目标最优化问题的一个非劣解。若再假设约束集合 R 为凸集，只要不断改变权系数 $\lambda_i(\lambda_i\geqslant 0)$，对其相应的加权和目标函数

$$U(x)=\sum_{i=1}^{m}\lambda_i f_i(x)$$

求出的最优解可以跑遍所有多目标问题

$$V\text{——}\max_{x\in R} f(x)$$

的非劣解集，但这方法只是从原则上(而且要有一定假设)可以求出所有非劣解，而在实际处理上却有一定困难。如何依次变动权系数，而使其得出最优解，正好得到所有非劣解，下面举例说明。

例 10 求 $V\text{——}\max\limits_{x\in R} f(x)$，其中

$$f(x)=\{f_1(x),f_2(x)\}^{\mathrm{T}}$$

$$f_1(x)=2x-x^2$$

$$f_2(x)=x$$

$$R=[0,2]$$

解 易看出这个多目标问题的非劣解

$$x^{*}\in[1,2]$$

而利用线性加权和方法，需要作新目标函数

$$U(x)=\lambda(2x-x^2)+(1-\lambda)x,\lambda\in[0,1]$$

在 R 中找其最优解，对 $U(x)$ 求导，并令其为零

$$U'(x)=\lambda(2-2x)+(1-\lambda)=0$$

可得

$$x=\frac{1+\lambda}{2\lambda}$$

显然当 $\lambda=1/3$ 时，$x^*=2$；$\lambda=1$ 时，$x^*=1$。当 λ 从 1/3 变到 1 时，即可得到全部非劣解 $[1,2]$；而 λ 从 0 变到 1/3 时，$U(x)$ 的最优解都是 $x^*=2$，即这时得不到新的非劣解。

17.6 多目标线性规划的解法

当所有目标函数是线性函数，约束条件也都是线性时，可有些特殊的解法。特别是泽勒内(Zeleny)等将解线性规划的单纯形法给予适当修正后，用来解多目标线性规划问题，或把多目标线性规划问题化成单目标的线性规划问题后求解，以下介绍两种方法。

17.6.1 逐步法(STEM)

逐步法是一种迭代法。在求解过程中，每进行一步，分析者把计算结果告诉决策者，决策者对计算结果作出评价。若认为已满意了，则迭代停止；否则分析者再根据决策者的意见进行修改和再计算，如此直到求得决策者认为满意的解为止，故称此法为逐步进行法或对话式方法。

设有 k 个目标的线性规划问题。

$$V\text{——}\max_{x\in R}\boldsymbol{C}x$$

其中 $R=\{x\,|\,\boldsymbol{A}x\leqslant b,x\geqslant 0\}$，$\boldsymbol{A}$ 为 $m\times n$ 矩阵

$\boldsymbol{C}$ 为 $k\times n$ 矩阵，也可表示为

$$\boldsymbol{C}=\begin{pmatrix}c^1\\ \vdots\\ c^k\end{pmatrix}=\begin{pmatrix}c_1^1,c_2^1,\cdots,c_n^1\\ \vdots\quad\vdots\quad\ddots\quad\vdots\\ c_1^k,c_2^k,\cdots,c_n^k\end{pmatrix}$$

求解的计算步骤为：

第一步：分别求 k 个单目标线性规划问题的解。

$$\max_{x\in R}c^jx,\quad j=1,2,\cdots,k$$

得到最优解 $x^{(j)}$，$j=1,2,\cdots,k$，及相应的 $c^jx^{(j)}$。显然

$$c^jx^{(j)}=\max_{x\in R}c^jx$$

并作表 $Z=(z_i^j)$，如表 17-2 所示。其中 $z_i^j=c^ix^{(j)}$，$z_j^j=\max\limits_{x\in R}c^jx=c^jx^{(j)}=M_j$

表 17-2

	z_1	z_2	z_i	z_k
$x^{(1)}$	z_1^1	z_2^1	$\cdots z_i^1 \cdots$	z_k^1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x^{(i)}$	z_1^i	z_2^i	$\cdots z_i^i \cdots$	z_k^i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x^{(k)}$	z_1^k	z_2^k	$\cdots z_i^k \cdots$	z_k^k
M_j	z_1^1	z_2^2	z_i^i	z_k^k

第二步：求权系数。

从表 17-2 中得到

$$M_j \text{ 及 } m_j = \min_{1 \leqslant i \leqslant k} z_i^j, \quad j=1,2,\cdots,k$$

为了找出目标值的相对偏差以及消除不同目标值的量纲不同的问题，进行如下处理。

$$\text{当} M_j > 0, \quad \alpha_i = \frac{M_j - m_j}{m_j} \frac{1}{\sqrt{\sum_{i=1}^{n} (c_i^j)^2}}$$

$$M_j < 0, \quad \alpha_j = \frac{m_j - M_j}{M_j} \frac{1}{\sqrt{\sum_{i=1}^{n} (c_i^j)^2}}$$

经归一化后，得权系数

$$\pi_j = \frac{\alpha_j}{\sum_{j=1}^{k} \alpha_j}, \quad 0 \leqslant \pi_j \leqslant 1, \quad \sum \pi_j = 1, \quad j = 1,2,\cdots,k$$

第三步：构造以下线性规划问题，并求解。

$$Lp(1)\begin{cases} \min \lambda \\ \lambda \geqslant (M_i - c^i x)\pi_i, \quad i=1,2,\cdots,k \\ x \in R, \quad \lambda \geqslant 0 \end{cases}$$

假定求得的解为$\bar{x}^{(1)}$，相应的 k 个目标值为 $c^1\bar{x}^{(1)}, c^2\bar{x}^{(1)}, \cdots, c^k\bar{x}^{(1)}$，若 $x^{(1)}$ 为决策者的理想解，其相应的 k 个目标值为 $c^1x^{(1)}, \cdots, c^kx^{(1)}$。这时决策者将$\bar{x}^{(1)}$的目标值进行比较后，认为满意了就可以停止计算。若认为相差太远，则考虑适当修正。如考虑对第 j 个目标宽容一下，即让点步，减少或增加一个 Δc^j，并将约束集 R 改为

$$R^1\begin{cases} c^j x \geqslant c^j \bar{x}^{(1)} - \Delta c^j \\ c^i x \geqslant c^i \bar{x}^{(1)} \qquad i \neq j \\ x \in R \end{cases}$$

并令第 j 个目标的权系数 $\pi_j = 0$，这表示降低这个目标的要求。再求解以下线性规划问题。

$$Lp(2)\begin{cases}\min \lambda \\ \lambda \geqslant (M_i - c^i x)\pi_i & i=1,2,\cdots,k, \quad i \neq j \\ x \in R^1, \lambda \geqslant 0\end{cases}$$

若求得的解为$\bar{x}^{(2)}$，再与决策者对话，如此重复，直到决策者满意为止。

例 11 试求解多目标线性规划问题。

$$\max z_1 = 100x_1 + 90x_2 + 80x_3 + 70x_4$$

$$\min z_2 = 3x_2 + 2x_4$$

$$R\begin{cases}x_1 + x_2 \geqslant 30 \\ x_3 + x_4 \geqslant 30 \\ 3x_1 + 2x_3 \leqslant 120 \\ 3x_2 + 2x_4 \leqslant 48 \\ x_i \geqslant 0, \quad i = 1 \sim 4\end{cases}$$

解 为了使问题的目标函数统一为求最大化的规划问题，将 z_2 化为

$$\max w_2 = -3x_2 - 2x_4$$

第一步：求理想解。

分别求解两个单目标线性规划问题

$$\max_{x \in R} z_1 \text{ 和 } \max_{x \in R} w_2$$

得到最优解 $x^{(1)} = (14, 16, 39, 0)^T$

相应的目标值 $z_1^1 = 5\,960, w_2^1 = -48$，即 $z_2^1 = 48$

$$x^{(2)} = (20, 10, 30, 0)^T$$

相应的目标值 $z_1^2 = 5\,300, w_2^2 = -30$，即 $z_2^2 = 30$

第二步：作 z 和求权系数表，如表 17-3 所示。

表 17-3

	z_1	w_2
$x^{(1)}$	5 960	−48
$x^{(2)}$	5 300	−30
M_j	5 960	−30

用表中 z 的数据，可计算得到

$$\alpha_1 = 0.000\,645, \quad \alpha_2 = 0.166\,4$$

于是求得权系数

$$\pi_1 = 0.003\,87, \quad \pi_2 = 0.996\,13$$

第三步：求解以下线性规划问题。

$$Lp(1)\begin{cases}\min \lambda \\ \lambda \geqslant 0.003\,87[5\,960-(100x_1+90x_2+80x_3+70x_4)] \\ \lambda \geqslant 0.996\,13(3x_2+2x_4-30) \\ x\in R,\quad \lambda\geqslant 0\end{cases}$$

由此可求得解(整数近似值)为

$$\bar{x}^{(1)}=(19,11,31,0)^{\mathrm{T}}$$

相应的目标函数值$\bar{z}_1^1=5\,370, w_2^1=-33$

第四步：对话再计算。

分析者把计算的结果告诉决策者，决策者将这结果与理想值$(Z_1^1,Z_2^2)=(5\,960,30)$进行比较，认为求得的 $\bar{Z}_2^1=33$ 已接近理想值 $Z_2^2=30$，而 $\bar{Z}_1^1=5\,370$，低于理想值 5 960 太多。决策者要求提高 Z_1 的值，为此他提出将 Z_2 提高到 36，以便使 Z_1 增大。这时分析者根据决策者的要求，将原约束条件修改为 R^1。

$$R^1\begin{cases}C^2x\leqslant 36\\ C^1x\geqslant 5\,370\\ x\in R\end{cases}$$

因将第二个目标值的要求放宽了，故权系数 $\pi_2=0$，于是有线性规划问题：

$$LP(2)\begin{cases}\min\lambda\\ \lambda\geqslant 5\,960-(100x_1+90x_2+80x_3+70x_4)\\ x\in R^1\end{cases}$$

求解 $LP(2)$得到

$$\bar{x}^{(2)}=(18,12,33,0)^{\mathrm{T}}$$

相应的目标值

$$\bar{Z}_1^{(2)}=5\,520,\quad \bar{Z}_2^{(2)}=36$$

若这时决策者对此结果表示满意，即停止计算。

17.6.2 妥协约束法

设有两个目标的情况，即 $k=2$

$$V-\max_{x\in R}\boldsymbol{C}x$$

$R=\{x\mid \boldsymbol{A}x\leqslant b, x\geqslant 0\}$，$\boldsymbol{A}$ 为 $m\times n$ 矩阵，$x\in E^n$，$b\in E^m$

$$\boldsymbol{C}=\begin{pmatrix}c^1\\ c^2\end{pmatrix}=\begin{bmatrix}c_1^1\cdots c_n^1\\ c_1^2\cdots c_n^2\end{bmatrix}$$

这方法的中心是引进一个新的超目标函数 $z=w_1c^1x+w_2c^2x$。w_1、w_2 为权系数，$w_1+w_2=1, w_i\geqslant 0, i=1,2$；此外构造一个妥协约束

$$R^1: w_1[c^1x - z_1^1] - w_2[c^2x - z_2^2] = 0 \quad x \in R$$

z_1^1、z_2^2 分别为 c^1x、c^2x 的最大值(当 $x \in R$)。求解的具体步骤为

第一步：解线性规划问题。

$$\max_{x \in R} c^1 x$$

得到最优解 $x^{(1)}$ 及相应的目标函数值 z_1^1。

第二步：解线性规划问题。

$$\max_{x \in R} c^2 x$$

得到最优解 $x^{(2)}$ 及相应的目标函数值 z_2^2。

在具体求解时可以先用 $x^{(1)}$ 试一试，看是否是 $\max_{x \in R} c^2 x$ 的最优解。若是，则这问题已找到完全最优解，停止求解；若不是，则求 $x^{(2)}$ 及相应的 z_2^2。

第三步：解下面三个线性规划问题之一。

$$\max_{x \in R^1} z, \quad \max_{x \in R^1} c^1 x, \quad \max_{x \in R^1} c^2 x$$

得到的解为妥协解。

例 12　试求解多目标线性规划问题。

$$\max \begin{cases} z_1 = 3x_1 + x_2 \\ z_2 = x_1 + 2x_2 \end{cases} \qquad R \begin{cases} x_1 + x_2 \leqslant 7 \\ x_1 \leqslant 5 \\ x_2 \leqslant 5 \\ x_1,\ x_2 \geqslant 0 \end{cases}$$

解　分别求解线性规划问题。

$$\max_{x \in R} z_1 \quad 和 \quad \max_{x \in R} z_2$$

得到最优解 $x^{(1)} = (5,2)$，$z_1 = 17$，$x^{(2)} = (2,5)$，$z_2 = 12$，见图 17-13。

若取 $w_1 = w_2 = 0.5$，表示等权妥协。则有超目标函数

$$z = 0.5[3x_1 + x_2] + 0.5[x_1 + 2x_2] = 2x_1 + 1.5x_2$$

妥协约束 R^1

$$0.5[3x_1 + x_2 - 17] - 0.5[x_1 + 2x_2 - 12] = 0$$

即 $\qquad x_1 - 0.5x_2 = 2.5, \quad x \in R$

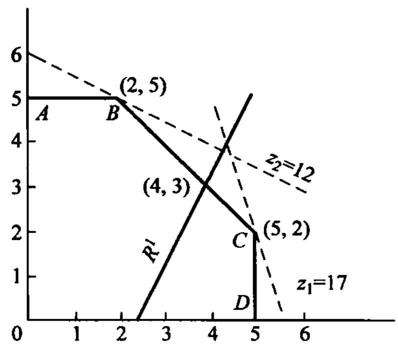


图　17-13

于是可以求得妥协解 $\bar{x} = (4,3)$，w_1、w_2 的取值可由决策者决定，这时可有不同的解。

解多目标线性规划问题的方法，还有线性目标规划法(详见本书的第 5 章)和其他方法。读者可参考有关文献资料。

17.7 层次分析法

层次分析法(analytic hierarchy process,AHP 法)是美国运筹学家沙旦(T. L. Saaty)于 20 世纪 70 年代提出的,是一种定性与定量分析相结合的多目标决策分析方法。特别是将决策者的经验判断给予量化,对目标(因素)结构复杂且缺乏必要的数据情况下更为实用,所以近几年来此法在我国实际应用中发展较快。

17.7.1 AHP 法原理

例如某工厂在扩大企业自主权后,有一笔企业留成的利润,这时厂领导要合理使用这笔资金。根据各方面反映和意见,提出可供领导决策的方案有:①作为奖金发给职工;②扩建职工食堂、托儿所;③开办职工业余技术学校和培训班;④建立图书馆;⑤引进新技术扩大生产规模等。领导在决策时,要考虑到调动职工劳动生产积极性,提高职工文化技术水平,改善职工物质文化生活状况等方面。对这些方案的优劣性进行评价,排队后,才能作出决策。

面对这类复杂的决策问题,处理的方法是,先对问题所涉及的因素进行分类,然后构造一个各因素之间相互联结的层次结构模型。因素分类:一为目标类,如合理使用今年企业留利××万元,以促进企业发展;二为准则类,这是衡量目标能否实现的标准,如调动职工劳动积极性,提高企业的生产技术水平;三为措施类,是指实现目标的方案、方法、手段等,如发奖金,扩建集体福利设施,引进新技术等。按目标到措施的自上而下地将各类因素之间的直接影响关系排列于不同层次,并构成一层次结构图,如图 17-14 所示。

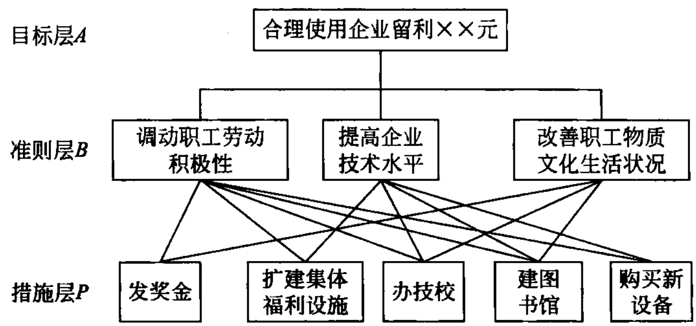


图 17-14

构造好各类问题的层次结构图是一项细致的分析工作,要有一定经验。根据层次结构图确定每一层的各因素的相对重要性的权数,直至计算出措施层各方案的相对权数。这就给出了各方案的优劣次序,以便供领导决策。

这个方法的原理是这样的。

设有 n 件物体 $A_1, A_2, \cdots, A_n$；它们的重量分别为 $w_1, w_2, \cdots, w_n$。若将它们两两的比较重量，其比值可构成 $n \times n$ 矩阵 $\boldsymbol{A}$。

$$\boldsymbol{A}=\begin{pmatrix} w_1/w_1 & w_1/w_2 & \cdots & w_1/w_n \\ w_2/w_1 & w_2/w_2 & \cdots & w_2/w_n \\ \vdots & \vdots & \ddots & \vdots \\ w_n/w_1 & w_n/w_2 & \cdots & w_n/w_n \end{pmatrix}$$

$\boldsymbol{A}$ 矩阵具有如下性质：若用重量向量

$$\boldsymbol{W}=(w_1, w_2, \cdots, w_n)^{\mathrm{T}}$$

右乘 $\boldsymbol{A}$ 矩阵，得到

$$\boldsymbol{AW}=\begin{pmatrix} w_1/w_1 & w_1/w_2 & \cdots & w_1/w_n \\ w_2/w_1 & w_2/w_2 & \cdots & w_2/w_n \\ \vdots & \vdots & \ddots & \vdots \\ w_n/w_1 & w_n/w_2 & \cdots & w_n/w_n \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} = n \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} = n\boldsymbol{W}$$

即
$$(\boldsymbol{A}-n\boldsymbol{I})\boldsymbol{W}=0$$

由矩阵理论可知，$\boldsymbol{W}$ 为特征向量，n 为特征值。若 $\boldsymbol{W}$ 为未知时，则可根据决策者对物体之间两两相比的关系，主观作出比值的判断，或用 Delphi 法来确定这些比值，使 $\boldsymbol{A}$ 矩阵为已知，故判断矩阵记作 $\bar{\boldsymbol{A}}$。

根据正矩阵的理论，可以证明：若 $\boldsymbol{A}$ 矩阵有以下特点（设 $a_{ij}=w_i/w_j$）：

(1) $a_{ii}=1$

(2) $a_{ij}=1/a_{ji}$，　$(i, j=1, 2, \cdots, n)$

(3) $a_{kj}=a_{ij}/a_{ik}$，　$(i, j=1, 2, \cdots, n)$

则该矩阵具有唯一非零的最大特征值 $\lambda_{\max}$，且 $\lambda_{\max}=n$。

若给出的判断矩阵 $\bar{\boldsymbol{A}}$ 具有上述特征，则该矩阵具有完全一致性。然而人们对复杂事物的各因素，采用两两比较时，不可能做到判断的完全一致性，而存在估计误差，这必然导致特征值及特征向量也有偏差。这时问题由 $\boldsymbol{AW}=n\boldsymbol{W}$ 变成 $\bar{\boldsymbol{A}}\boldsymbol{W}'=\lambda_{\max}\boldsymbol{W}'$，这里 $\lambda_{\max}$ 是矩阵 $\bar{\boldsymbol{A}}$ 的最大特征值，$\boldsymbol{W}'$ 便是带有偏差的相对权重向量。这就是由判断不相容而引起的误差。为了避免误差太大，所以要衡量 $\bar{\boldsymbol{A}}$ 矩阵的一致性，当 $\boldsymbol{A}$ 矩阵完全一致时，因 $a_{ii}=1$，

$$\sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} a_{ii} = n$$

存在唯一的非零 $\lambda=\lambda_{\max}=n$。而当 $\bar{\boldsymbol{A}}$ 矩阵存在判别不一致时，一般是 $\lambda_{\max} \geqslant n$。这时

$$\lambda_{\max} + \sum_{i \neq \max} \lambda_i = \sum_{i=1}^{n} a_{ii} = n$$

由于是
$$\lambda_{\max} - n = -\sum_{i \neq \max} \lambda_i$$

以其平均值作为检验判断矩阵一致性指标（CI）

$$\mathrm{CI}=\frac{\lambda_{\max}-n}{n-1}=\frac{-\sum\limits_{i\neq\max}\lambda_i}{n-1}$$

当 $\lambda_{\max}=n$,CI=0,为完全一致;CI 值越大,判断矩阵的完全一致性越差。一般只要 CI≤0.1,认为判断矩阵的一致性可以接受,否则重新进行两两比较判断。

判断矩阵的维数 n 越大,判断的一致性将越差,故应放宽对高维判断矩阵一致性的要求。于是引入修正值 RI,见表 17-4,并取更为合理的 CR 为衡量判断矩阵一致性的指标。

$$\mathrm{CR}=\frac{\mathrm{CI}}{\mathrm{RI}}$$

表 17-4

维数	1	2	3	4	5	6	7	8	9
RI	0.00	0.00	0.58	0.90	1.12	1.24	1.32	1.41	1.45

17.7.2 标度

为了使各因素之间进行两两比较得到量化的判断矩阵,引入 1～9 的标度。根据心理学家的研究提出:人们区分信息等级的极限能力为 7±2,特制定表 17-5。

因为自己与自己比是同等重要的,因此对角线上元素不用作判断比较,只需要给出矩阵对角线上三角形中的元素。

可见 $n\times n$ 矩阵,只需要给出 $\frac{n(n-1)}{2}$ 个判断数值。

除表 17-5 的标度方法以外,还可以用其他标度方法。

表 17-5

标　度 a_{ij}	定　　义
1	i 因素与 j 因素同样重要
3	i 因素比 j 因素略重要
5	i 因素比 j 因素较重要
7	i 因素比 j 因素非常重要
9	i 因素比 j 因素绝对重要
2,4,6,8	为以上两判断之间的中间状态对应的标度值
倒数	若 j 因素与 i 因素比较,得到判断值为 $a_{ji}=1/a_{ij}$,$a_{ii}=1$

17.7.3 层次模型

根据具体问题一般分为目标层、准则层和措施层(见图 17-15)。复杂的问题可分为总目标层、子目标层、准则层(或制约因素层)、方案措施层,或分为层次更多的结构(见

图 17-16)。下面举若干例子加以说明。

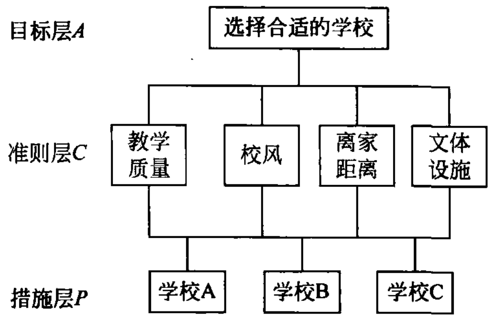


图 17-15

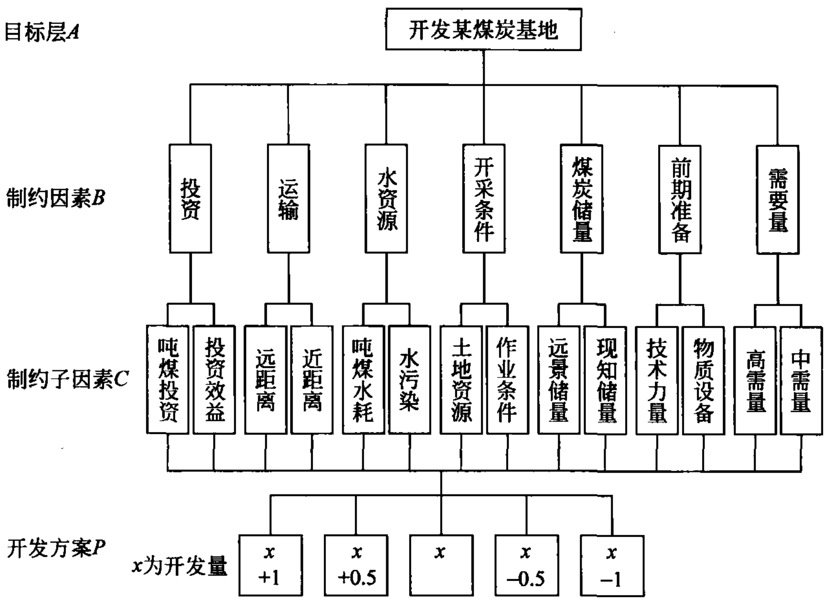


图 17-16

按给出的层次结构模型,设为目标层 A、准则层 C(有 k 个准则因素)、措施层 p(有 n 个方案)。由决策者用其他方法给出各层因素之间的两两比较得出 A-C 判断矩阵为

A	C_1	C_2	$\cdots$	C_k
C_1	a_{11}	a_{12}	$\cdots$	a_{1k}
C_2	a_{21}	a_{22}	$\cdots$	a_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
C_k	a_{k1}	a_{k2}	$\cdots$	a_{kk}

然后分别给出 C_i-P 的判断矩阵($i=1,2,\cdots,k$)。

C_i	P_1	P_2	$\cdots$	P_n
P_1	a_{11}	a_{12}	$\cdots$	a_{1n}
P_2	a_{21}	a_{22}	$\cdots$	a_{2n}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
P_n	a_{n1}	a_{n2}	$\cdots$	a_{nn}

用近似法计算各判断矩阵的最大特征值和特征向量。

17.7.4 计算方法

一般地讲,在 AHP 法中计算判断矩阵的最大特征值与特征向量,并不需要很高的精度,故用近似法计算即可。最简单的方法是求和法及其改进的方法,但方根法更好,这里只介绍方根法。

1. 方根法

这是一种近似计算法,其计算步骤为:

(1) 计算判断矩阵每行所有元素的几何平均值

$$\bar{\omega}_i=\sqrt[n]{\prod_{j=1}^{n}a_{ij}} \qquad i=1,2,\cdots,n$$

得到 $\bar{\omega}=(\bar{\omega}_1,\bar{\omega}_2,\cdots,\bar{\omega}_n)^{\mathrm{T}}$。

(2) 将 $\bar{\omega}_i$ 归一化,即计算

$$\omega_i=\frac{\bar{\omega}_i}{\sum\limits_{i=1}^{n}\bar{\omega}_i} \qquad i=1,2,\cdots,n$$

得到 $\bar{\omega}=(\omega_1,\omega_2,\cdots,\omega_n)^{\mathrm{T}}$,即为所求特征向量的近似值,这也是各因素的相对权重。

(3) 计算判断矩阵的最大特征值 $\lambda_{\max}$

$$\lambda_{\max}=\sum_{i=1}^{n}\frac{(A\bar{\omega})_i}{n\bar{\omega}_i}$$

其中 $(A\bar{\omega})_i$ 为向量 $A\omega$ 的第 i 个元素。

(4) 计算判断矩阵一致性指标,检验其一致性。

当各层次的诸因素的相对权重都得到后,进行措施层的组合权重计算。

2. 组合权重计算

设有目标层 A、准则层 C、方案层 P 构成的层次模型(当层次更多的模型,计算相同),目标层 A 对准则层 C 的相对权重为

$$\bar{\omega}^{(1)}=(\omega_1^{(1)},\omega_2^{(1)},\cdots,\omega_k^{(1)})^{\mathrm{T}}$$

准则层的各准则 C_i,对方案层 P n 个方案的相对权重为

$$\bar{\omega}_l^{(2)}=(\omega_{1l}^{(2)},\omega_{2l}^{(2)},\cdots,\omega_{nl}^{(2)})^{\mathrm{T}} \qquad l=1,2,\cdots,k$$

那么各方案对目标而言,其相对权重是通过权重 $\bar{\omega}^{(1)}$ 与 $\bar{\omega}_l^{(2)}$ ($l=1,2,\cdots,k$)组合而得到的,

其计算可采用表格式进行(见表 17-6)。

这时得到的 $V^{(2)}=(v_1^{(2)},v_2^{(2)},\cdots,v_n^{(2)})^T$ 为 P 层各方案的相对权重。

表 17-6

P层	因素及权重 C层 C_1 $C_2\cdots C_k$ $\omega_1^{(1)}\ \omega_2^{(1)}\cdots\omega_k^{(1)}$	组合权重 $V^{(2)}$
P_1	$\omega_{11}^{(2)}\ \omega_{12}^{(2)}\cdots\omega_{1k}^{(2)}$	$v_1^{(2)}=\sum_{j=1}^{k}\omega_j^{(1)}\omega_{1j}^{(2)}$
P_2	$\omega_{21}^{(2)}\ \omega_{22}^{(2)}\cdots\omega_{2k}^{(2)}$	$v_2^{(2)}=\sum_{j=1}^{k}\omega_j^{(1)}\omega_{2j}^{(2)}$
⋮	⋮	⋮
P_n	$\omega_{n1}^{(2)}\ \omega_{n2}^{(2)}\cdots\omega_{nk}^{(2)}$	$v_n^{(2)}=\sum_{j=1}^{k}\omega_j^{(1)}\omega_{nj}^{(2)}$

注 记

决策论从单目标发展到多目标是在理论和实践上的一个飞跃。用多目标决策方法来处理决策问题更能满足实践的要求。从理论上看多目标决策方法吸取了行为科学等社会科学方面的成就，强调了决策者本人在决策过程中的中心地位；此外在解的概念方面，提出有效解(非劣解)、满意解、妥协解等更能反映复杂现实情况的解的概念。在应用方面具有广阔的前景。1974 年 Zeleny 所编著的《线性多目标规划》一书较系统地总结了各种处理线性多目标规划问题的方法。20 世纪 70 年代中期后，各种交互式的多目标决策方法迅速发展起来。1982 年 Zeleny 编写了第一本论述多目标决策问题的教科书。20 世纪 70 年代以来多目标决策发展的另一特点是各种计算方法都已有较实用的软件，便于实用。我国自 1977 年以来也开始了多目标决策方面的研究和应用工作，为广大的运筹学工作者和实际工作部门所重视，并取得了一定的成果。

参考资料

[1] Zeleny M. Multiple Criteria Decision Making. New York, McGraw-Hill Book Company, 1975

[2] Churchman C W. Introduction to Operations Research. John Wiley and Son, 1957

[3] Салуквадзе М Е. Задачи векторной оптимизации в теории управления. медниереба, 1975

[4] Hwang C L and Masud A S. Multiple Objective Decision Making. Methods and Applications. Springer—Verlag. Berlin, 1979

[5] Hwang C L and Yoon K S. Multiple Attribute Decision Making. Springer Verlag, 1981

[6] 陈珽. 决策分析. 北京：科学出版社，1987

[7] 顾基发，魏权龄. 多目标决策问题. 应用数学与计算数学，1980. 1

Part 11

第 11 篇

启发式方法

在前面各章中，讨论了一些常用的和标准的运筹学模型，给出并证明了它们的优化算法，借助这些算法，可以求出模型的最优解。在论述时，大多数是从科学分析的角度出发，尽可能给出理论的解释和数学的证明，力求既能解决实际问题，又有逻辑的严密性。

但必须说明的是，在解决问题时所采用的大多数模型并不是自然而然地呈现在研究者或管理者的面前，而是需要通过人们的感知去认识和发现问题，经过分析然后提炼成模型。而且所发现的问题往往是嵌入在大系统中的某个子系统或子子系统，它们关联着社会、经济、文化和生态环境等诸多方面。此外很多模型往往是在边界清晰、条件明确、判别准则合理等较理想的条件下才能进行求解的，所得结果与实际情况常常存在差别，有时差别可能很大。

人们对客观世界的一般认识，或对某一具体对象的认识，都要经过由暗到明和由浅入深的过程，大致包含以下几个阶段：

(1) 水平 1：朦胧。在复杂的社会经济系统中，存在着很多未被人们认识清楚，但又能若明若暗地感觉得到的问题的胚芽，这类对象是否需要进一步研究，尚不够清楚或还有争议，需进一步认真分析。

(2) 水平 2：提出问题。在人们对研究对象进行系统观察和分析的基础上，将问题的胚芽加以总结抽象，而形成明确的问题。

(3) 水平 3：建立数学模型。通过对问题的系统研究和深刻理解，明确其本质属性及与环境的关系，将问题抽象并一般化，进而表述为数学模型。运用数学模型不仅能解决一个具体问题，而且可以用来分析和处理一类问题，或便于研究某问题在多参数、多方案情况下的性态及其优化。

(4) 水平 4：求解数学模型。通常是运用某些已有的算法或提出新的算法(包括移植、改造、综合)，通过一定的步骤，求出问题的解。人们往往希望得到问题的最优解或满意解。

(5) 水平 5：使用和改进。将其用于解决实际问题并在实践中验证和改进模型及其算法。这是提高人们认识水平、有效解决实际问题和发展科学不可缺少的步骤。

运筹学工作者或管理工作者面对的实际问题是多种多样的，有的问题的结构与运筹学中已有模型十分相近，可用相应的标准运筹学模型和算法来解决；有的问题与运筹学已有模型的结构有较大的差异，这时就很难直接套用运筹学中已有的方法来求解。为能有效解决这种问题，就需要学会去发现和创造能解决更复杂实际问题的新思路、新模型和新算法。

当人们面对一些新的难题需要解决时，一种自然的想法就是看它和以往处理过的问题有无相似或相通之处，能否从以往的经验(包括自身的和别人的)中得到某种借鉴，或从社会发展和自然规律中得到某种启示，从而为解决新问题找到一些可行的思路和方法。事实上，寻求并接受启发，是人们认识问题、分析问题和解决问题的一种常用思维模式，是实现不断进步和超越的一种可行途径。有意识和主动地寻求启发常会加快解决问题的速度和提高解决问题的质量，达到事半功倍的效果。

CHAPTER 18
第18章*

启发式方法

18.1 基本概念

18.1.1 问题的结构

有些实际问题的结构比较清晰，各元素之间的关系明确，边界清楚，容易为人们所认识，能够通过建模和使用一定的算法求得解决，这类问题称为**良好结构**问题。一般而言，良好结构问题具有以下特征：

(1) 能建立起正确反映该问题性质的一种"**可接受**"模型，与问题有关的主要信息可纳入模型之中。

(2) 模型所需要的数据能够获得。

(3) 模型可解，能拟定出求解的程序性步骤和求解方法，而且，得到的解能体现解决问题的可行方案。

(4) 可拟定出明确的准则，用以判定解的可行性和最优性。

(5) 求解所需的**计算量**不太大，所需的费用不太多。

对于良好结构问题，常可用传统的(标准的)运筹学方法加以解决。如果问题的结构不良，使用传统的运筹学方法去处理就难以奏效。这时，与其歪曲事实，忽略或修正某些重要的条件，勉强使用某种标准模型而使问题易于求解，还不如保持问题的本来面目，建立基本符合问题实际情况的非标准模型。前者虽可用已有的标准算法求解，但由于问题的模型失真，得到的解难于符合实际和付诸实施；后者由于模型涉及因素多，结构复杂，而与传统的标准模型相去甚远，难于套用已有的标准算法。在后面这种情况下，为得到近似可用的解，分析人员必须运用自己的感知和洞察力，从与其有关而较基本的模型及算法中寻求其间的联系，从中得到启发，去发现适于解决该问题的思路和途径，这种方法称为**启发式方法**(heuristic method)，由此建立的算法称为**启发式算法**(heuristic algorithm)。

18.1.2 启发式方法的特点

由上面的叙述可知，启发式方法是寻求解决问题的一种较好的方法和策略；当然，它也可以是面向某种具体问题的一种求解方法。它建立在人们经验和判断的基础之上，体

现了人的主观能动作用和创造力。

用启发式方法解决问题时强调“满意”,常常是得到满意解,决策者就认为可以了,而不去刻意追求最优性和探求最优解。之所以这样,其原因是:

(1) 很多问题不存在严格最优解(例如目标之间矛盾的多目标问题),这时,对目标的满意性要求常比最优性追求更能准确地描述人们的选择行为。

(2) 对有些问题,要得到它的最优解所花的代价太大,难以实现,也不合算。

(3) 从实际决策的需要出发,有时要求解具有过高的精度没有必要。

假定为解决某类问题设计了一个算法,它能用于求解所有这类问题,而且获得最优解的计算工作量可表示为这类问题“大小”的多项式函数,就称这个算法是确定型的多项式算法,简称为**多项式算法**或**有效算法**。很多**组合优化**问题(如设施定位问题、旅行售货员问题、多个工件在多个设备上的加工排序问题等)不存在多项式算法,要求其最优解就需花费巨大的代价。

用启发式方法求解问题常常是通过迭代过程实现的,因而需要拟定出一套科学合理的解的搜索规则。为能得到满意解,在整个迭代过程中要及时掌握解的变化情况,不断注意和吸收新的信息,及时考察所使用的求解策略,必要时改变原来拟定的不合适的或过时的策略,建立新的搜索规则,注意从失败中吸取教训,并逐步缩小搜索范围。

在工业、商业、管理、社会等方面的很多问题,目前不可能找到多项式算法,为使问题得以解决,自然需要求助于启发式方法。在构建启发式方法时,为提高其效果,人们常注意到:①多种相关知识的集成和综合运用;②多种方法的合理结合;③系统状态在多时段中的良性动态演化;④考虑各种干扰因素的影响;⑤顾及多个相关主体利益的均衡。这样做虽然增加了构建和运用启发式方法的难度,但却使它更贴近实际并易于达到更好的效果。

启发式方法具有下述优点:

(1) 计算步骤简单,要求的理论基础不高,可由未经高级训练的人员实现;

(2) 比优化方法常可减少大量的计算工作量,从而显著节约开支和时间;

(3) 易于将定量分析与定性分析相结合。

使用启发式方法时应注意得到的解的质量,使由于采用启发式方法而使最终决策效果明显有所改善。在选用方法时要考虑是否有现成的标准优化方法可以采用,如果使用优化方法的工作量可以接受,则应慎重考虑是否要选用启发式算法。

18.1.3 启发式策略

用启发式方法解决问题时,需要采用一定的策略。下面列出几个常用的策略,使用时可根据问题的性质和要求选用其中之一;为得到理想的效果,也可将几个策略联合起来使用。

(1) 问题构建策略。在对提出的问题认真研究和全面分析的基础上,明确研究目标和欲解决的关键问题,由此建立的模型应如实和正确反映问题的性质、条件、要求、结构和

目标，防止问题扭曲和表述失真。在满足表述正确和重点明确的前提下，力求简单、明晰和便于解决。

(2) 逐步构解策略。一个完整的解通常是由若干个分量组成的。当用该策略时，应建立某种规则，按一定次序每次确定解的一个分量，直至得到包含所有解分量的一个完整的解为止。

(3) 分解合成策略。为求解一个复杂的大问题，可首先将其分解为若干个小的子问题，再选用合适的方法(包括启发式方法、优化方法、模拟方法等)按一定顺序求解每个子问题，根据子问题之间及其与总问题的关系(例如递阶关系、包含(嵌套)关系、平行关系等)，将子问题的解作为下一阶子问题的输入，或在相容原则下将子问题的解进行综合，经合成最后得到总问题合乎要求的解。

(4) 改进策略。运用这一策略时，首先从一个初始解(初始解不必一定是可行解)出发，然后对解的质量(包括它产生的目标函数值、可行性及可接受性等)进行评价，并采用某种启发式方法设计改进规则，对解加以改进，反复进行如上的评价和改进，直至得到满意的解为止。

为获得初始解，可用逐步构解策略或(和)分解合成策略，也可使用其他近似方法。在启发式方法中，好的初始解常可大大提高求解效率和质量。

(5) 搜索学习策略。本策略包括在解空间中的定向搜索以及在搜索过程中发现和收集新的信息，并根据对新信息的分析，重新确认或改变搜索方向，修正搜索参数，消去不必要的搜索范围，以有效提高搜索效率，尽快获得问题的解。

近年来发展起来的贪婪算法(greedy algorithm)、模拟退火(simulated annealing)、禁忌搜索(tabu search)、遗传算法(genetic algorithm)、人工神经网络(artificial neural network)等方法具有启发式方法的特性(有人称其为亚启发式方法)，它们为构建和运用具体问题的启发式方法提供了启示和方便。贪婪算法的每一步都以局部或当前利益为着眼点作出选择，这符合一个贪婪者的心理状态和行为准则，但由于其目光短浅，使以此作出的“最佳”选择最终可能达不到全局最优。模拟退火和禁忌搜索方法都是为了跳出局部最优而设计的，前者基于对热力学过程的模拟，属于一种随机性算法；后者属于一种确定性算法(当然也可以在其中增加一些概率因素)，它采用一种“记忆”设置，能记住已被搜索的解，从而禁止再对它们进行多余的搜索。遗传算法是一种并行、随机、自适应优化算法，它基于适者生存的规律，模拟自然演化过程，把搜索空间映射为遗传空间，每个解编码为一个向量(染色体)，将问题表达为染色体的适者生存过程，通过按其适应性的选择、交叉、变异等遗传操作，剔除适应性低的染色体，保留适应性高的染色体，如此反复迭代，直至满足预定的优化指标。人工神经网络是由神经元互连构成的网络，它利用仿生学和微电子技术，模拟人脑的智能活动，具有强大的学习功能，能从样例中学习并获取知识。

18.2 应用及例子

启发式方法在理论上是基于对比、分析、探索、综合的一种科学方法，同时它在使用上又是一种艺术，其成功程度取决于使用者的水平。为能较好地运用启发式方法解决实际问题，使用者必须具有比较广阔的知识，较好的基础和善于观察、分析、联想的能力，以及善于从类似问题的解决方法中获得启示的敏感性和素质。

下面结合几个例子研究如何使用启发式方法来解决实际问题。

18.2.1 多个工件在多个设备上加工的排序问题

n个工件在m台设备上加工的最优顺序问题，目前尚无多项式算法。此处为简单计，仅考虑两台设备A和B，研究n个工件($j=1,2,\cdots,n$)在这两台设备上顺次加工时应如何排列工件的顺序，才能使总加工时间(从在设备A上加工第一个工件起到在设备B上加工完最后一个工件止这段时间)尽可能短。此处要求每个工件都先在设备A上加工，加工完后再在设备B上加工。

在第10章中曾经指出，如果工件在设备A上的加工顺序与在设备B上的加工顺序不同，由于增加了等待时间，将使总加工时间延长。因此，在研究该问题时可将这种情况排除在外，不予考虑。即使如此，可能的排序方案仍有$n!$个，随着工件数n的增多，其计算工作量增加很快。下面寻求用启发式方法的解决途径。

下面这个例子给出了6个工件分别在设备A和设备B上的加工时间A_j和B_j(分钟)(见表18-1)，所有工件均先在设备A上加工，再在设备B上加工。要求确定使总加工时间最短的工件加工顺序。

表18-1 加工时间表

设备	工件					
	1	2	3	4	5	6
A	30	60	60	20	80	90
B	70	70	50	60	30	40

为了得到该问题的启发式方法，此处运用逐步构解策略。先考虑工件1和工件2，其可能的排序方案有两个：1-2和2-1(图18-1)。由于$B_1=B_2$，$A_1<A_2$，故将工件1排在前面加工所需的总加工时间较少。再看工件2和工件3，由于$A_2=A_3$，$B_3<B_2$，故将工件3排在工件2的后面加工所需的总加工时间较少(图18-2)。

虽然图18-1和图18-2仅分别比较了两个工件的不同加工顺序，且依据的是例子中给定的特定情况，但我们可由此得到启发，将其推广应用到n个工件在两台设备上的加工顺序问题，拟定出以下启发式迭代步骤。

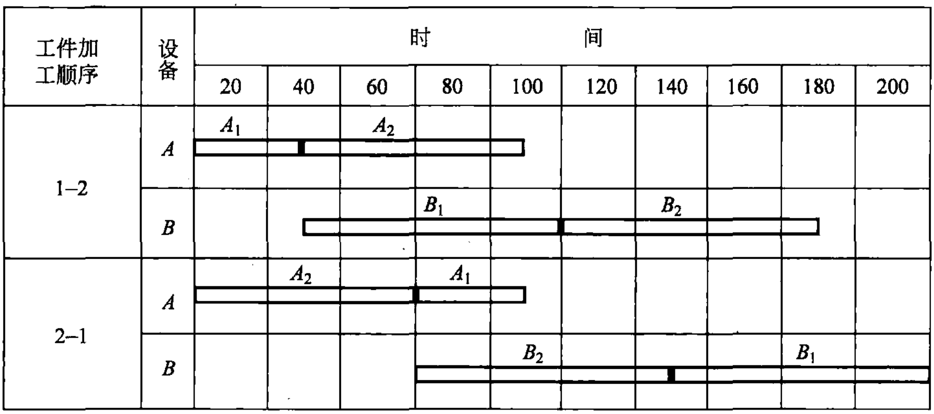


图 18-1

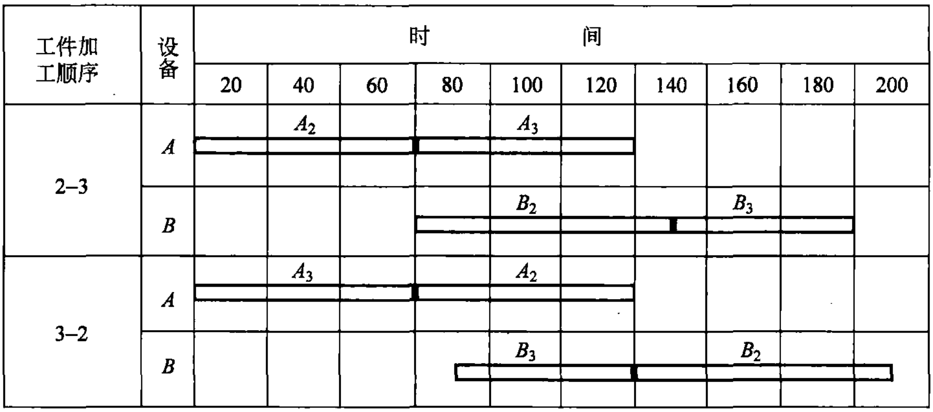


图 18-2

(1) 令 $i=1,k=0$。

(2) 找最小加工时间，即

$$t_r=\min\{A_1,A_2,\cdots,A_n,B_1,B_2,\cdots,B_n\} \tag{18-1}$$

(3) 若 $t_r=A_j$，则安排工件 j 为第 i 个加工工件，并置 $i:=i+1$；若 $t_r=B_j$，则安排工件 j 为第 $(n-k)$ 个加工工件，并置 $k:=k+1$。

(4) 将 A_j 和 B_j 从式(18-1)的工件加工时间表中删去，即不再考虑已排好加工顺序的工件 j。

(5) 转步骤(2)，直至式(18-1)中的工件加工时间表变成空集。

现将上述步骤应用于表 18-1 所示的排序问题，得到各工件的加工顺序如下：

$$4\to1\to2\to3\to6\to5$$

总加工时间等于 370 分钟，具体情况示于图 18-3 中。

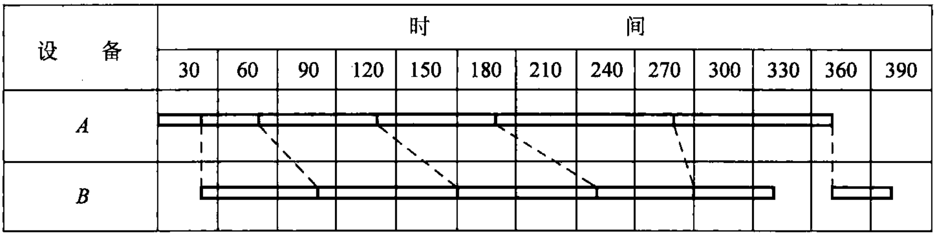


图 18-3

需要指出的是，对在两台设备上加工 n 个工件的问题来说，用图 18-3 所示方法求得的解是最优解。但是，如将其扩展应用到在 m 台设备上加工 n 个工件的一般加工排序问题，所得结果一般就不再是最优解了。然而，用这种思想却常常能得到较好的解。

18.2.2 旅行售货员(旅行商)问题

旅行售货员问题(traveling salesman problem，TSP)指的是：一个售货员从某城市出发，访问 n 个城市(售货)各一次且仅一次，然后回到原地，他走什么样的路线才能使走过的路程最短(或旅行费用最低)。在第 11 章曾经说到，这个问题就是寻求总权最小的哈密尔顿(Hamilton)回路问题。到目前为止，一般 TSP 问题还没有多项式算法，对于较大的问题(例如 n 大于 40)就需要使用启发式方法求解。

下面介绍两种可用的启发式算法。

1. *C-W* 节约算法

这种算法由 Clarke 和 Wright 提出，下面说明其基本思想和迭代步骤。

假定有 n 个访问地(例如城市)，把每个访问地看成一个点，并取其中的一个点为基点(起点)，例如以 1 点为基点。首先将每个点与该基点相连接，构成线路 $1\to j\to 1(j=2,3,\cdots,n)$，这样就得到了一个具有 $n-1$ 条线路的图(当然，这时尚未形成 Hamilton 回路)。旅行者按此线路访问这 n 个点所走的路程总和为

$$z=2\sum_{j=2}^{n}c_{1j} \tag{18-2}$$

其中 c_{1j} 为由点 1 到点 $j(j=2,\cdots,n)$ 的路段长度，注意此处假定 $c_{1j}=c_{j1}$(对所有 j)。

若连接点 i 和点 $j(i,j\neq 1)$，即使旅行者走弧(i,j)时(当然这时就不再经过弧$(i,1)$和$(1,j)$)，所引起的路程节约值 $s(i,j)$ 可计算如下

$$\begin{aligned} s(i,j)&=2c_{1i}+2c_{1j}-(c_{1i}+c_{1j}+c_{ij})\\ &=c_{1i}+c_{1j}-c_{ij} \end{aligned} \tag{18-3}$$

对不同的点对(i,j)，$s(i,j)$越大，旅行者通过弧(i,j)时所节约的路程越多，因而应优先将其安排到旅行线路中去，使旅行者旅行时通过这一条弧。

在具体应用该方法时，可按以下步骤进行。

(1) 选取基点，例如选取点 1 为基点。将基点与其他各点连接，得到 $n-1$ 条线路 $1\to j\to 1(j=2,3,\cdots,n)$。

(2) 对不违背限制条件的所有可连接点对 (i,j)，如下计算其节约值（i,j 不为基点）

$$s(i,j)=c_{1i}+c_{1j}-c_{ij}$$

(3) 将所有 $s(i,j)$ 按其值的大小由大到小排列。

(4) 按 $s(i,j)$ 的上述顺序，逐个考查其端点 i 和 j，若满足以下条件，就将弧 (i,j) 插入到线路中。其条件是：

① 点 i 和点 j 不在一条线路上；

② 点 i 和点 j 均与基点相邻。

(5) 返回步骤(4)，直至考查完所有可插入弧 (i,j) 为止。

通过以上各迭代步骤，使问题的解逐步得到改善，最后达到满意解（也有可能达到最优解）。

例 1 试用 C-W 节约算法求解下述旅行售货员问题，已知有 7 个访问点，其位置如图 18-4 中所示。

解 先按图 18-4 给出的数据计算各点对之间的欧氏距离 $c(i,j)$，计算结果列入距离表（表 18-2）中。由于已假设 $c_{ij}=c_{ji}$（对所有 i 和 j），故该表中各元素的值以主对角线为对称。

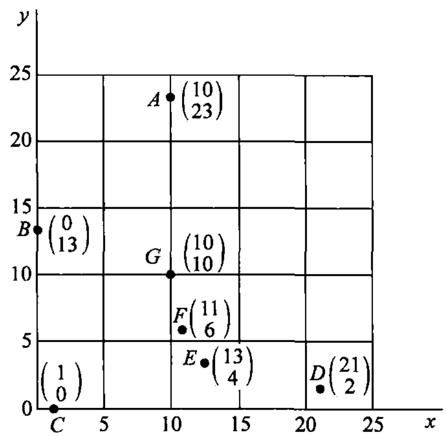


图 18-4

表 18-2

始点	终点						
	A	B	C	D	E	F	G
A	0	14.14	24.70	23.71	19.24	17.03	13.00
B	14.14	0	13.04	23.71	15.81	13.04	10.44
C	24.70	13.04	0	20.10	12.65	11.66	13.45
D	23.71	23.71	20.10	0	8.25	10.77	13.60
E	19.24	15.81	12.65	8.25	0	2.83	6.71
F	17.03	13.04	11.66	10.77	2.83	0	4.12
G	13.00	10.44	13.45	13.60	6.71	4.12	0

取 G 为基点（也可取其他的访问点为基点），构成初始线路图（图 18-5）。再用式(18-3)计算将弧 $(i,j)(i,j\neq G)$ 插入到线路中时引起的路程节约值，并按节约值由大到小的顺序将它们填入表 18-3 中。

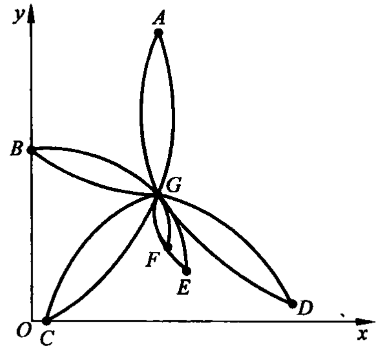


图 18-5

表 18-3

序号	弧	节约值	序号	弧	节约值
1	(D,E)	12.06	9	(A,D)	2.89
2	(B,C)	10.85	10	(A,C)	1.75
3	(A,B)	9.30	11	(B,F)	1.52
4	(E,F)	8.00	12	(B,E)	1.34
5	(C,E)	7.51	13	(A,E)	0.47
6	(C,D)	6.95	14	(B,D)	0.33
7	(D,F)	6.95	15	(A,F)	0.09
8	(C,F)	5.91			

按节约值从大到小的顺序，对每条弧加以考查，看能否将其插入到旅行线路中。若能将其插入，就对旅行线路作相应的改变。本例的线路调整过程如表 18-4 所示。

表 18-4

序号	弧	线路与说明	插入该弧的节约值
0		G→A→G，G→B→G，G→C→G，G→D→G，G→E→G，G→F→G	
1	(D,E)	G→A→G，G→B→G，G→C→G，G→D→E→G，G→F→G	12.06
2	(B,C)	G→A→G，G→B→C→G，G→D→E→G，G→F→G	10.85
3	(A,B)	G→A→B→C→G，G→D→E→G，G→F→G	9.30
4	(E,F)	G→A→B→C→G，G→D→E→F→G	8.00
5	(C,E)	E 点与基点 G 不相邻，不插入	0
6	(C,D)	G→A→B→C→D→E→F→G	6.95

当插入弧(C,D)之后，线路已包含所有要访问的点，这时算法终止。用该方法得到的旅行线路（见图 18-6）是

$$G \to A \to B \to C \to D \to E \to F \to G$$

该条旅行线路的总长度

$$\begin{aligned} z &= 2(13.0+10.44+13.45+13.60+6.71+4.12)- \\ &\quad (12.06+10.85+9.30+8.00+6.95) \\ &= 2\times 61.32-47.16=75.48 \end{aligned}$$

用这种方法得出的解不一定是最优解，但这里得到的解是本例的最优解。

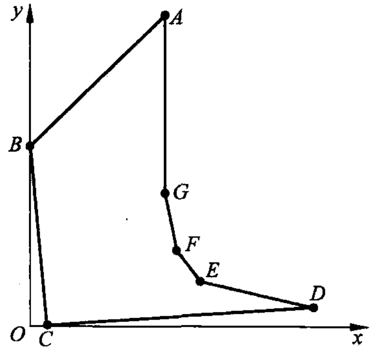


图 18-6

在实际问题中，点对间的距离应以路段的真实长度为准（不一定为两点间的欧氏距离），其间的往返距离也可以不相等。这时，插入某一段弧引起的路程节约值的计算应根据实际情况进行。

2. 几何法

该方法由 J. P. Norback 和 R. F. Love 提出，它基于对各访问点构成的几何图形的分析，以此确定初始旅行线路和不在初始线路上的各点的插入顺序和插入位置。

根据一般几何观察可知，最短旅行线路应具有以下直观性质：

① 线路自身不相交；

② 各段线路应处于由所有访问点形成的凸包上或其凸包内部（这里所说的凸包(convex hull)是指包含所有访问点的最小凸集）。

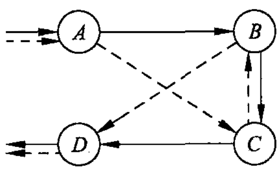


图 18-7

图 18-7 中给出了连接同样 4 个点的两条线路，显然，自相交的线路（图 18-7 中虚线所示者）$A\to C\to B\to D$ 要比不自交的线路 $A\to B\to C\to D$ 为长。

根据上述分析，启发我们拟定出求解旅行售货员问题的下述迭代步骤：

(1) 找出由欲访问各点构成的凸包。

(2) 在凸包上的点，按其出现的自然顺序访问（注意不要使旅行线路自交），从而形成一初始旅行线路。

(3) 将不在初始旅行线路上的各个点 I（位于凸包内的访问点），与已在旅行线路上的所有点相连。设 P 与 Q 为已在旅行线路上的任两个相邻点，$\angle P_0I_0Q_0$ 为所有$\angle PIQ$ 角度中的最大者，则将 I_0 插入到 P_0 和 Q_0 之间。

(4) 重复进行步骤(3)，每次在旅行线路上增加一个新点，直至所有欲访问点都被引入到旅行线路中为止。这时就构成了一条哈密尔顿回路。

下面用这种方法求解例 1。其迭代过程示于图 18-8 中。开始时构成凸包 $ABCDA$（图 18-8(a)），以它为初始旅行线路，然后将不在初始旅行线路上的 E、F 和 G 三点分别与 A,B,C,D 四点相连（图 18-8(b)），考查以 E,F 和 G 为角顶，分别以 AB,BC,CD 和

DA 为对边形成的各个角度，由于$\angle CED$ 最大，故将点 E 插入在 C 和 D 两点之间，形成新的旅行线路 $ABCEDA$(图 18-8(c))。现不在访问线路上的点为 F 和 G，连接 EF 和 EG，考查以 F 和 G 为角顶的各角，因$\angle DGA$ 最大，从而将 G 点插入到 D 点和 A 点之间，这时的访问线路变为 $ABCEDGA$(图 18-8(d))。如上继续，将点 F 插入到 D 点和 G 点之间，这就得到了本问题的哈密尔顿回路，可以它作为本问题的解(图 18-8(e))，其线路总长为

$$14.14+13.04+12.65+8.25+10.77+4.12+13.00=75.97$$

作为比较，将该问题的最优解示于图 18-8(f)中，其最优线路总长度为 75.48，它可以通过交换线路中 E 和 D 这两个相邻点的连接顺序而得到。

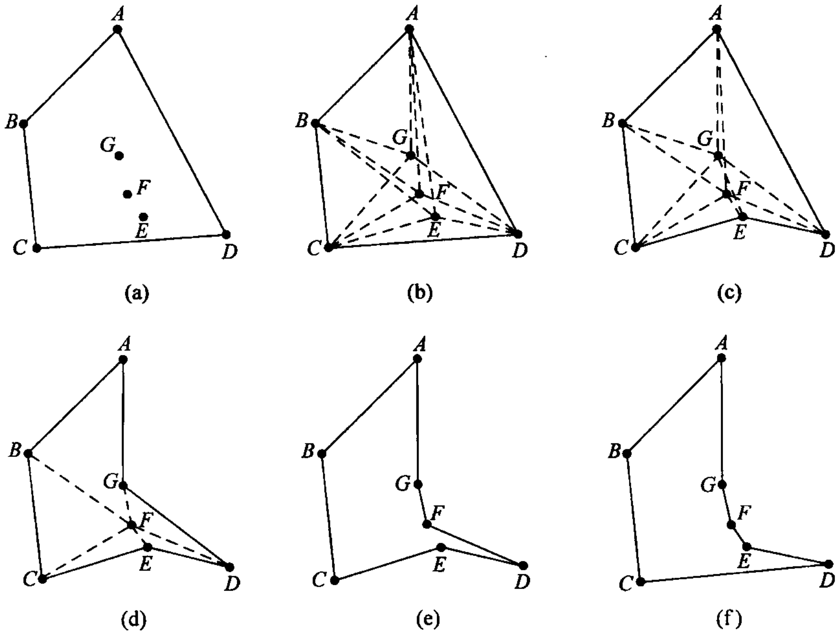


图 18-8

用几何法虽未得到例 1 的最优解，但它的精确度还是很高的，常常可以达到比较满意的结果。

除上述两种方法外，还有一些其他的算法可用于求解旅行售货员问题，在解决实际问题时，可同时使用几种算法，从中选取最好的结果。

18.2.3 车辆调度问题

车辆调度问题(vehicle scheduling problem，VSP)是由 Dantzig 和 Ramser 于 1959 年提出的，虽经多人潜心研究，但由于其复杂性大，目前仍未找到多项式算法，专家们多把精力集中于研究高质量的启发式算法方面。

所谓 VSP 问题，一般指的是：对一系列发货点和收货点，组织适当的行车路线，使车

辆有序地通过它们，在满足一定的约束条件下(例如货物需求量与发送量，交发货时间，车量容量限制，行驶里程限制，行驶时间限制等)，力争实现一定的目标(如空驶里程最短，运输费用极小，车辆按时到达，使用车辆数量尽可能少等)。

车辆调度问题的分类法很多，例如可根据车辆满载与否分为满载问题与非满载问题，根据可用车场数分为单车场问题与多车场问题，根据可用车辆的车型数分为单车型问题与多车型问题，根据决策者的要求分为单目标问题与多目标问题等。

下面研究单车型、多车场、满载运输问题的一种启发式算法，并考虑使总空驶里程极小这一目标(它对运输公司的运输效益有极大影响)。

1. 数学模型

设某运输公司有 n 个车场可以使用，即可从车场 $A_{m+1},A_{m+2},\cdots,A_{m+n}$ 发出空车和接收空车，它们与各货运业务的发货点和收货点位于同一个道路网上。各车场可派出的空车数分别为 $b_{m+1},b_{m+2},\cdots,b_{m+n}$，可接收的空车数分别为 $b'_{m+1},b'_{m+2},\cdots,b'_{m+n}$。假定运输公司要完成的货运业务有 m 项：$A_1,A_2,\cdots,A_m$，其货运量分别是 $g_1,g_2,\cdots,g_m$，完成各项业务所需的车辆数分别为 $a_1,a_2,\cdots,a_m$。

按照每项业务的要求将货物由发货点运到收货点全部为重车行驶，其行车路线可由网络上任意点对间的最短路方法确定。

现考虑第 i 项货运业务，用 i' 表示其发点，i'' 表示其收点。由于由 i' 到 i'' 是重车行驶，按照运输计划的要求，不管选用什么运输组织方案，都必须完成此项工作，并已假定选择的运输路径为“最短路径”，故在研究使总空驶里程极小化问题时，可将这一运输业务看成一个点，称为**收缩点**或**重载点** i。

对于每一个重载点 i，为运出其货物量 g_i 需 a_i 辆空车，它们将货物运抵目的地卸车后，又提供出 a_i 辆空车，这些空车驶向其他货运业务的发点(这时为空车行驶)，继续装货执行运输任务。由此可见，执行运输任务的每辆货车都如此交替地进行空驶和重驶，直至完成一天(或半天)的运输任务，返回某一车场为止。

设由点 i 发往点 j (i,j 为车场或重载点)的空车数为 x_{ij}，其空驶里程为 c_{ij}，则使总空驶里程极小化的空车调度问题可描述如下(称为问题 T)。

$$
\text{T:}\begin{cases}
\min z=\displaystyle\sum_{i=1}^{m+n}\sum_{j=1}^{m+n}c_{ij}x_{ij}\\
\displaystyle\sum_{j=1}^{m+n}x_{ij}=a_i, & i=1,2,\cdots,m\\
\displaystyle\sum_{j=1}^{m+n}x_{ij}\leqslant b_i, & i=m+1,m+2,\cdots,m+n\\
\displaystyle\sum_{i=1}^{m+n}x_{ij}=a_j, & j=1,2,\cdots,m\\
\displaystyle\sum_{i=1}^{m+n}x_{ij}\leqslant b'_j, & j=m+1,m+2,\cdots,m+n\\
x_{ij}\geqslant 0 \text{ 且为整数}
\end{cases}
\tag{18-4}
$$

此处 $a_i(a_j)$ 可由下式得出

$$\begin{cases} a_i = g_i/q, & \text{若 } g_i/q \text{ 为整数} \\ a_i = \left[\dfrac{g_i}{q}\right]+1, & \text{若 } g_i/q \text{ 不为整数} \end{cases} \tag{18-5}$$

式中：$\left[\dfrac{g_i}{q}\right]$为数值不大于 g_i/q 的最大整数；

q 为一辆车的可载量。

c_{ij} 根据实际情况选用，可取为点 i 到点 j 的广义最短距离。为避免由车场发出的空车不经重载点直接驶向车场，令

$$c_{ij}=M, \qquad i,j=m+1,m+2,\cdots,m+n$$

其中 M 为一足够大的正数。

问题 T 的运输表示于表 18-5。

表 18-5

发空车 \ 收空车		重载点 A_1 A_2 $\cdots$ A_m	车场 A_{m+1} A_{m+2} $\cdots$ A_{m+n}	发车数
重载点	A_1 A_2 $\vdots$ A_m	$C—C$	$C—F$	a_1 a_2 $\vdots$ a_m
车场	A_{m+1} A_{m+2} $\vdots$ A_{m+n}	$F—C$	$F—F$	b_{m+1} b_{m+2} $\vdots$ b_{m+n}
收车数		a_1 a_2 $\cdots$ a_m	b'_{m+1} b'_{m+2} $\cdots$ b'_{m+n}	

2. 算法

众所周知，运输问题的表上作业法是一种有效的标准算法，而且能方便地得出整数最优解，在构造求解问题式(18-4)的启发式算法时，应注意充分利用表上作业法的优点。

首先仅考虑重载点，得问题 T_0

$$\begin{cases} \min z_1 = \sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij}x_{ij} \\ \sum_{j=1}^{m} x_{ij}=a_i, \qquad i=1,2,\cdots,m \\ \sum_{i=1}^{m} x_{ij}=a_j, \qquad j=1,2,\cdots,m \\ x_{ij}\geqslant 0 \text{ 且为整数} \end{cases} \tag{18-6}$$

这是一个一般的产销平衡运输问题，可直接用表上作业法求解，设求出的最优解为 $X^{(0)}=(x_{ij}^{(0)})$。以这个解中的非零变量的值作为问题 T 中对应变量的值，其他变量取值为零，这就得到了问题 T 的一个可行解。

但是，用上述方法得到的解无法以它为根据进行派车，因而是不可接受的。为此，需要设计一套解的判别和调整规则，使从 $X^{(0)}$ 出发，经有限步迭代，而得到问题 T 的可接受最优解或满意解。其步骤如下。

(1) 解的扩展

对解 $X^{(0)}$ 中的每一个非零分量 $x_{ij}^{(0)}>0 \quad (i,j=1,2,\cdots,m)$，计算

$$\delta_{ij}=\min_{m+1\leqslant i\leqslant m+n}\{c_{ij}\mid\bar{b}_i>0\}+\min_{m+1\leqslant j\leqslant m+n}\{c_{ij}\mid\bar{b}'_j>0\}-c_{ij} \tag{18-7}$$

式中，

$$\bar{b}_i=b_i-\sum_{j=1}^{m}x_{ij}^{(0)},\quad i=m+1,\cdots,m+n \tag{18-8}$$

$$\bar{b}'_j=b'_j-\sum_{i=1}^{m}x_{ij}^{(0)},\quad j=m+1,\cdots,m+n \tag{18-9}$$

由式(18-7)算出的 δ_{ij} 值，为按下述方法调整 $x_{ij}^{(0)}$ 一个单位引起的空驶里程增加量。若由式(18-7)得出的 δ_{ij} 来自 k 行和 l 列$[k,l\in(m+1,m+n)]$，则把 $x_{ij}^{(0)}>0$ 扩展至 x_{kj} 和 x_{il}，这时，将 $x_{ij}^{(0)}$、$x_{kj}^{(0)}$ 和 $x_{il}^{(0)}$ 三者的值分别调整为

$$\begin{cases}x_{ij}:=x_{ij}^{(0)}-\min\{x_{ij}^{(0)},\bar{b}_k,\bar{b}'_l\}\\x_{kj}:=x_{kj}^{(0)}+\min\{x_{ij}^{(0)},\bar{b}_k,\bar{b}'_l\}\\x_{il}:=x_{il}^{(0)}+\min\{x_{ij}^{(0)},\bar{b}_k,\bar{b}'_l\}\end{cases} \tag{18-10}$$

解的扩展工作按 δ_{ij} 的大小由小到大依次进行，直至找出要求的可接受解，即在表 18-5中的 F—C 和 C—F 区含有适当的非零解分量。如此得到的解记为 $X^{(1)}=(x_{ij}^{(1)})$，显然，它对问题 T 是可行的。当 $\delta_{ij}<0$ 时，按此调整得到的解 $X^{(1)}$ 优于 $X^{(0)}$。

(2) 解的收缩

本步骤是步骤(1)的逆过程。当表 18-5 中 F—C 区和 C—F 区的非零解分量的值 $\left(\sum_{j=1}^{m}\sum_{k=m+1}^{m+n}x_{kj}^{(1)}\text{ 和 }\sum_{i=1}^{m}\sum_{l=m+1}^{m+n}x_{il}^{(1)}\right)$比派车数要大时，需将非零解分量向 C—C 区收缩。这时，对每一对 $x_{kj}^{(1)}>0$ 和 $x_{il}^{(1)}>0[k,l\in(m+1,m+n);\ i,j\in(1,m)]$，计算

$$\delta'_{ij}=\min\{c_{ij}-c_{kj}-c_{il}\mid x_{kj}^{(1)}>0,x_{il}^{(1)}>0\} \tag{18-11}$$

并以此进行解的调整

$$\begin{cases}x_{kj}:=x_{kj}^{(1)}-\min\{x_{kj}^{(1)},x_{il}^{(1)}\}\\x_{il}:=x_{il}^{(1)}-\min\{x_{kj}^{(1)},x_{il}^{(1)}\}\\x_{ij}:=x_{ij}^{(1)}+\min\{x_{kj}^{(1)},x_{il}^{(1)}\}\end{cases} \tag{18-12}$$

调整后得到的解记为 $X^{(2)}=(x_{ij}^{(2)})$，这个解也是问题 T 的可行解。

3. 安排行车路线

经上述调整得到的解($X^{(1)}$ 和 $X^{(2)}$)，是问题 T 的可接受可行解，可作为安排行车路线

的依据。

安排行车路线时首先在可接受可行解 X 的非零分量中寻求下述序列

$$x_{k_1k_2}>0,\quad x_{k_2k_3}>0,\quad x_{k_3k_4}>0,\quad \cdots,\quad x_{k_pk_q}>0 \tag{18-13}$$

其中下标 $k_1,k_q\in[m+1,m+n]$，即 $x_{k_1k_2}$ 和 $x_{k_pk_q}$ 分别位于表 18-5 中的 $F—C$ 区和 $C—F$ 区。如此即可得到一条初始行车路线如下

$$\underset{\text{车场}}{A_{k_1}}\to\underbrace{A_{k_2}\to A_{k_3}\to A_{k_4}\to\cdots\to A_{k_p}}_{\text{重载点}}\to\underset{\text{车场}}{A_{k_q}}$$

有了初始行车路线(一条或若干条)之后，再根据具体的约束条件(例如一条线路的长度限制，时间限制，各条线路的均匀性要求等)进行调整(包括某些线路的截短或合并)，最后由调度员选择合适的驾驶员执行。

实用证明，本启发式方法的运算速度快，精度高，可达到很好的效果。需要指出，解的调整工作(扩展、收缩)应结合安排行车路线反复进行。

注　　记

由于现实问题的复杂性以及已有规范性方法的局限性，使“启发式方法”成为人们解决复杂的、非结构化问题的一种常用方法和策略。近年来，在启发式方法理论研究和应用技术不断深入和扩大的同时，人们对人工神经网络(artificial neural networks)算法、模拟退火(simulated annealing)算法、禁忌搜索(tabu search)算法和遗传算法(genetic algorithm)等亚启发式算法的研究和应用也得到了广泛重视和发展，取得了许多有价值的研究成果。对该方面有兴趣的读者，请参阅有关专著和教材。

[1]　[日]玄光男，程润伟著. 遗传算法与工程设计. 北京：科学出版社，2000

[2]　胡守仁，余少波，戴葵著. 神经网络导论. 北京：国防科技大学出版社，1993

[3]　陈森发编著. 复杂系统建模理论与方法. 南京：东南大学出版社，2005

启发式方法在计算机和信息科学方面的应用，可参阅：

[美]George E. Luger 著. 史忠植，张银奎，赵志崑译. 人工智能：复杂问题求解的结构和策略. 北京：机械工业出版社，2004

习题

18.1　什么是启发式方法？说明用启发式方法解决实际问题的过程和步骤。

18.2　在解决实际问题时应如何运用启发式策略？除本书上列出的几个启发式策略之外，你认为还有什么样的策略可以使用？

18.3　对在多台设备上加工多个工件的工件排序问题来说，应如何衡量不同排序方案的优劣？你认为应有哪些准则？这些准则的适用条件是什么？请举出两个实例加以详细说明。

18.4　试说明 C-W 节约算法的基本思想，你认为还可用它解决哪些方面的问题？举

例加以说明。

18.5 试将 Norback 和 Love 提出的几何法与 C-W 节约算法进行比较。

18.6 说明本书所述货运车辆优化调度算法的原理和求解步骤，并绘出求解过程框图。请简要回答以下问题：

(1) 若有两种车型的车可用，书中提出的模型应怎样修改？在书中所提算法的启发下，试拟定出一套求解的迭代步骤。

(2) 你认为应如何将书中提出的模型和算法推广到多目标的情形。

18.7 表 18-6 给出了 12 个工件在设备 A 和 B 上的加工时间，要求：

(1) 若所有工件都先在设备 A 上加工，再在设备 B 上加工，试确定使总加工时间最短的工件加工顺序，并计算总加工时间。

(2) 若工件 8～12 先在设备 B 上加工，再在设备 A 上加工，其他条件同上，试设计一启发式算法，以计算最小总加工时间和安排相应的工件最优加工顺序。(提示：可考虑将工件 1～7(要求先在设备 A 上加工，然后再在设备 B 上加工)和工件 8～12(要求先在设备 B 上加工，然后再在设备 A 上加工)这两组工件分别排序，再将二者按总加工时间尽量短的要求组合起来。)

表 18-6

设备	工件											
	1	2	3	4	5	6	7	8	9	10	11	12
A	5	8	11	2	4	7	12	3	9	3	6	10
B	5	9	4	3	7	6	9	4	5	8	9	4

18.8 有 4 个工件 J_1,J_2,J_3,J_4，要求在三台设备 A,B,C 上顺次加工，各工件在各设备上的加工时间示于表 18-7 中，试构造一启发式算法，用于寻求使总加工时间最短的工件加工顺序。

表 18-7

设备	工件			
	J_1	J_2	J_3	J_4
A	5	10	9	5
B	7	5	7	8
C	9	4	5	10

18.9 有 10 个城市，它们在坐标系中的位置如表 18-8 所示，试完成以下工作：

(1) 用 C-W 节约算法求出经过每个城市一次且仅一次的一条最短线路。

(2) 用 Norback 和 Love 提出的几何法，求出经过上述每个城市一次且仅一次的最短线路。

(3) 比较上述两种方法得出的结果，并设计一种启发式方法，对上述较差的结果进行改进。

表 18-8

坐标	城市									
	1	2	3	4	5	6	7	8	9	10
x	0	5	8	7	10	15	16	18	18	20
y	0	20	12	4	15	4	18	8	15	17

18.10 有一运输问题，它有 3 个重载点和 2 个车场，其运输表如表 18-9 所示。表中小方框内的数字为两点间的车辆空驶距离，1，2 和 3 三个重载点对应的三项运输业务的重载里程(已将装卸车时间折算在内)分别为 7，8 和 9，其他有关情况如表中所示。此外，要求车辆的每条行车路线总长度(包括重驶、空驶及装卸车所用时间的折算长度)L 在 45～60 之间。试用本章给出的车辆优化调度启发式算法，求出其满意的可接受可行解，并据此排出行车路线。

表 18-9

发空车		收空车					发车数
		重载点			车场		
		1	2	3	4	5	
重载点	1	4	12	10	8	10	10
	2	2	8	8	6	10	6
	3	12	4	6	14	2	7
车场	4	6	4	6	M	M	≤4
	5	14	2	8	M	M	≤6
收车数		10	6	7	≤8	≤7	

参考资料

[1] Kanti Swarup P K. Gupta and Man Moohan. Operations Research. Sultan Chand & Sons. 1982
[2] Hamdy A Taha. Operations Research: An Introduction. 3rd Edition. The Macmillan Company, 1980
[3] 郭耀煌等编著. 运筹学原理与方法. 成都: 西南交通大学出版社,1994
[4] 郭耀煌,李军. 满载问题的车辆路线安排. 系统工程学报,1995 年第 2 期
[5] 郭耀煌. 安排城市卡车行车路线的一种新算法. 系统工程学报,1989 年第 2 期
[6] Zbigniew Michalewicz and David B. Fogel,曹宏庆等译. 如何求解问题——现代启发式方法. 北京: 中国水利水电出版社,2003